October 2023

Leumit Synthetic Data

About MDClone

MDClone is a technology firm focused on unlocking healthcare data and empowering exploration, discovery, and collaboration to improve patients' health. Founded 2016 in Israel and headquartered in Beer Sheba, MDClone has partnered with major health systems, government entities, payers, and life science companies worldwide.

The MDClone ADAMS Platform includes innovative synthetic data publishing capabilities that further enhance the protection of patient privacy. Our unique technology enables any user of a healthcare organization to organize, access, and protect the privacy of patient data, empowering healthcare workers to transform ideas into actionable insights and take action rapidly.

The MDClone Solution

MDClone, in collaboration with Leumit, has prepared a set of synthetic data file(s) for a cohort of Leumit patients derived from real patient records using the MDClone Synthetic Data Lake (SDL) technology.

Leumit synthetic data is a synthetic version of Leumit original data pertinent to a cohort including the entire lifetime of the patient from birth to death/disengagement from the available Leumit health system data before June 2018.

The synthetic data enables access to all potential variables that could be used for a specific analysis while preventing any practical attempt for patient re-identification. It can be used for clinical data analysis and ML/AI model development.

Leumit Synthetic Data content

There are ~130K synthetic patients in the data set.
The dataset consists of variables for sixteen (16) clinical events that include demographics, encounters, visits, diagnoses, medications, measurements, lab results, procedures, and death.

Synthetic data events are provided in Comma Separated Value (CSV) format, with one file per clinical event that can be joined using the synthetic patient identifier.

See separate document with list of events and properties.

Leumit Synthetic Data privacy settings and limitations

The SDL is designed to provide maximum data utility while ensuring patient privacy. There are some preliminary mechanisms to protect the privacy that take place before the synthesis process itself:

1. Simple de-identification, which includes removal of addresses, names, email addresses, phone numbers, etc.
2. Rare conditions, medications and other categorical data elements appearing less than a predefined and small number of times in the original data lake are removed.
3. The synthesis is based on a random sample from the original cohort.
4. Data elements having hierarchical structures are mixed by deciding on a baseline hierarchy level to preserve, allowing the levels below it to mix between similar patients.
5. Dates, including years, are shifted. This can cause past or future dates to emerge with respect to the patients' timeline (e.g. even though we have limited the dates to 2003-2018, the synthetic data may show earlier or later years)
6. Numerical variables are mixed between similar patients using similarity metrics. This means that very unique patterns common to few patients will not show up in the synthetic data.

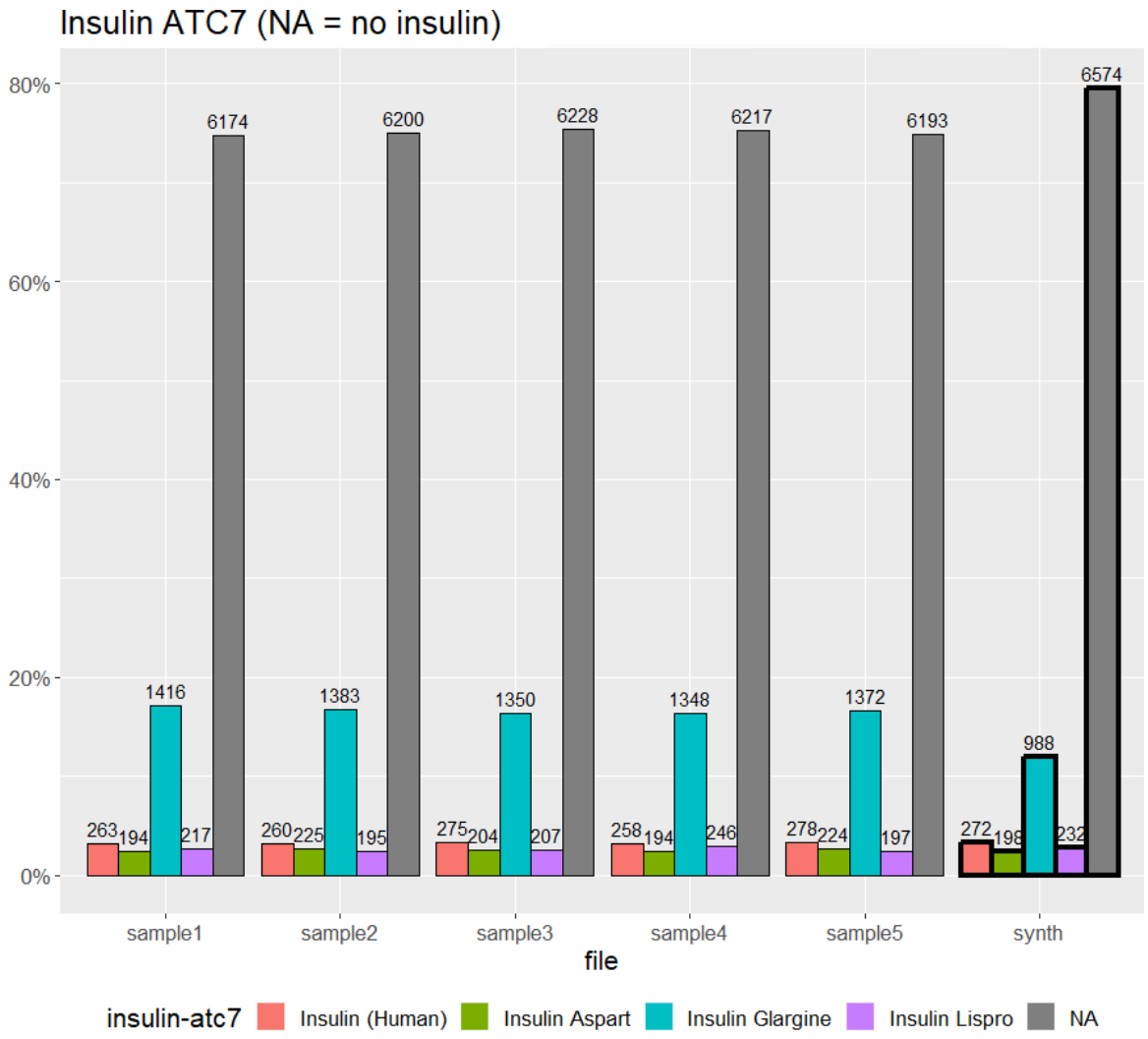
There are, therefore, certain limitations of using the synthetic data as a surrogate for original data:

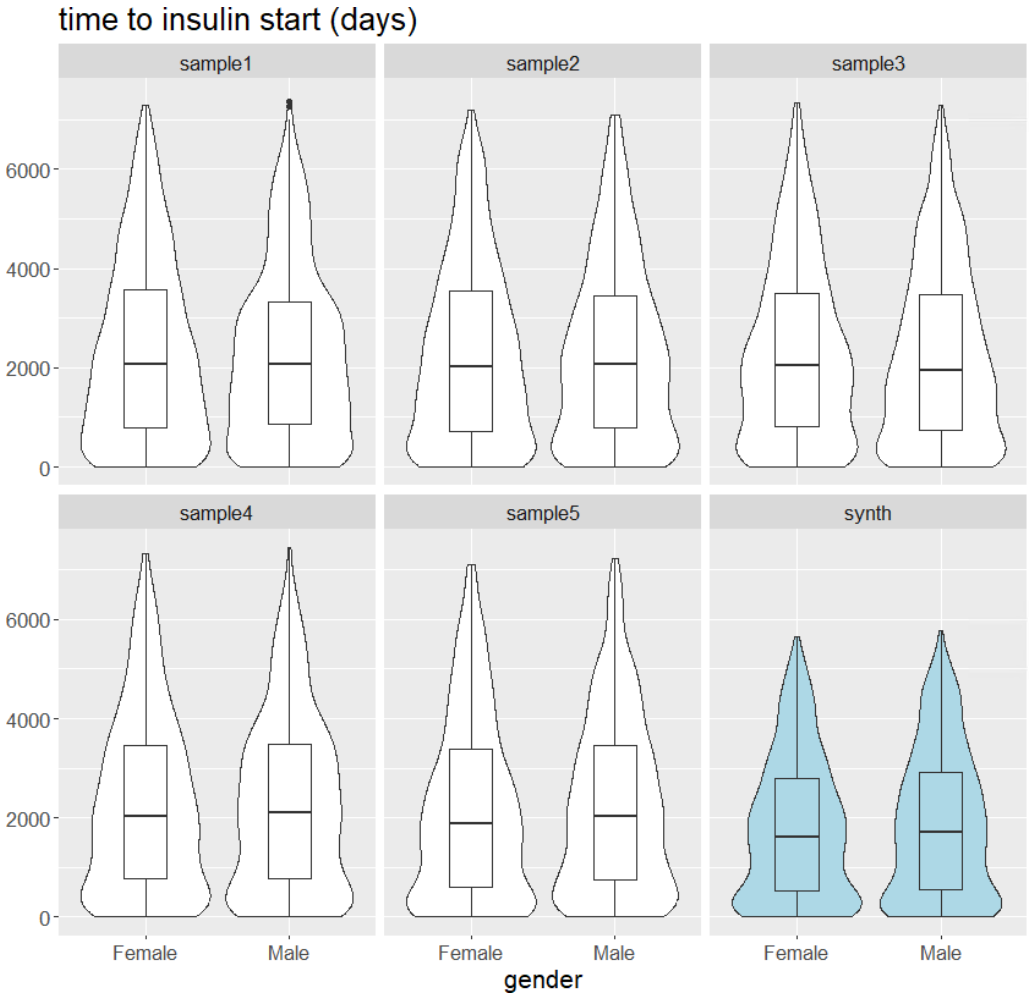
1. The exact number of original patients cannot be inferred.
2. Absolute date values are not generally preserved including days of week, season or year. However, time relationships are accurately captured (time between diagnosis and medication, admission length of stay, time to readmission, etc.).
3. Data elements that have hierarchies, like diagnoses and medications, will not be provided at the most granular level. For medications, the researcher will only be exposed to the active ingredient (ATC7). For diagnoses, we provide the ICD9 level.
4. **Patients will be represented well in the Synthetic data for Cohort size > 1000.**

Synthetic Data Evaluation

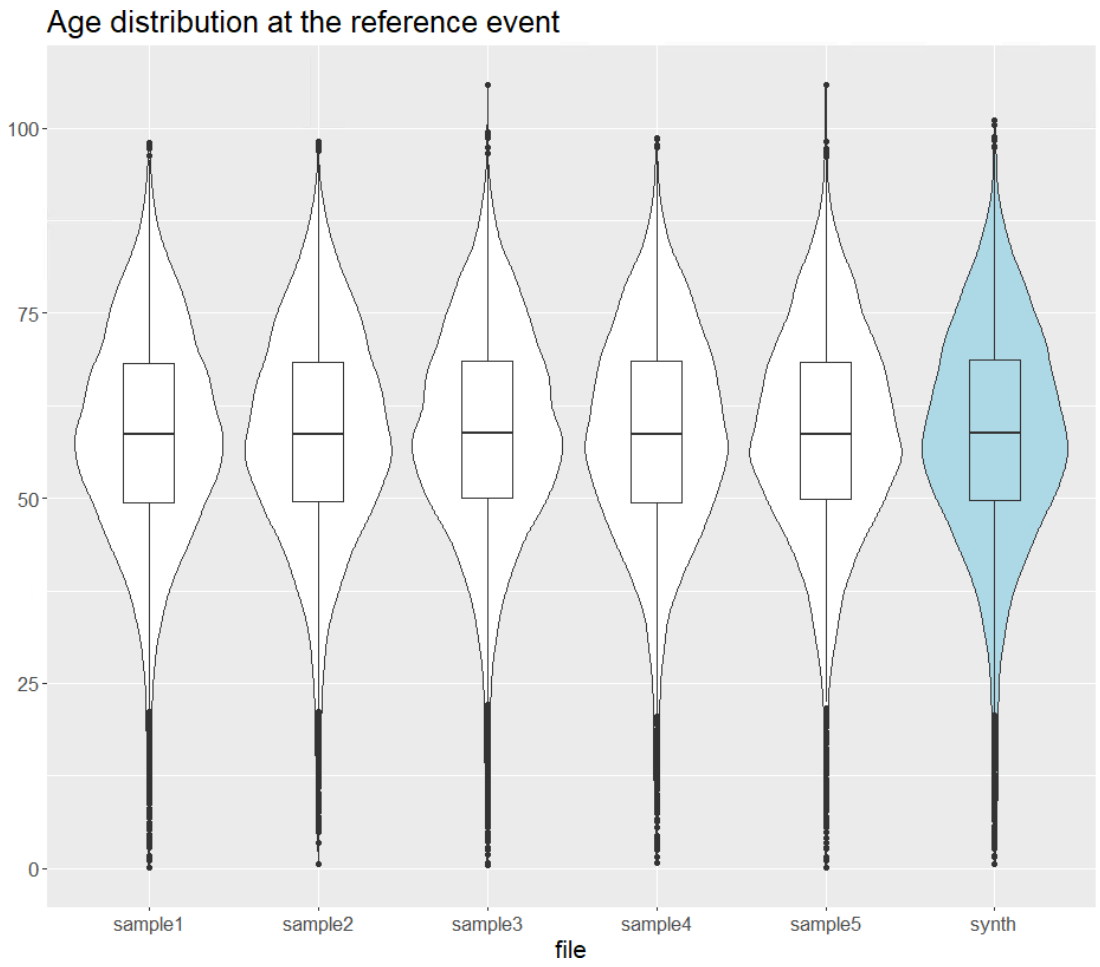
The original data was used as a reference and employed a combination of univariate, exploratory data analysis (EDA), multivariate, and Machine Learning modeling techniques to assess the quality of the synthetic data.
For example:

1. Comparing histograms of 5 original samples vs one synthetic sample of Insulin ATC7.
   The distribution looks similar.


Insulin ATC7 (NA = no insulin)
80%
60%
40%
20%
0%
6174
6200
6228
6217
6193
6574
1416
1383
1350
1348
1372
988
263
194
217
260
225
195
275
204
207
258
194
246
278
224
197
272
198
232
sample1
sample2
sample3
sample4
sample5
synth
file
insulin-atc7
Insulin (Human)
Insulin Aspart
Insulin Glargine
Insulin Lispro
NA



2. Comparing the distribution of patients who were diagnosed with diabetes type 1 or 2 before 2019 between 5 original samples and one synthetic sample.
   The distribution looks similar.


Age distribution at the reference event
100
75
50
25
0
sample1
sample2
sample3
sample4
sample5
synth
file