– Understanding and correcting missing and erroneous data in diabetes-related time series

The purpose of this challenge is to address missing and erroneous data from RWD, in particular diabetic patients, that consists of multiple time series.

Apply until **16 April 2023** / Xplorers Camp on **26 April 2023**

Question to be solved

Scientific Question: Build a framework/ concept of missing and erroneous data, or any other type of biased data, for diabetes related time series coming from RWE. Propose a methodology of how the effects of missing/ erroneous data can be mitigated.

General background

About the Challenge: At Roche we are at the forefront of deploying Digital Health strategies to enable utilization of RWE for building preventative care solutions in many areas, for example Diabetes care. Such RWE come with a number of challenges, for example missing or erroneous data. An example of RWE for diabetic patients consists of multiple time series -insulin, glucose and patient meals.

Personalized decision support models require the development of in-silico models from such data. Before using such data for modeling, SMEs spend significant time in curating the data. Improper curation results in model convergence issues and modeling biases.

Automated curation approaches will help SMEs save significant time, avoid onerous activities and enable rapid development of in-silico models.

Data types & technologies

- Time series data
- Predictive Modeling
- Machine Learning
- Exploratory data analysis

- Feature engineering
- Model explainability tools
- Data preprocessing and creating preprocessing pipelines
- Data engineering
- Real time analytics

Supporting material or links

- https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
- Multiple Imputation for Incomplete Data in Epidemiologic Studies | American Journal of Epidemiology | Oxford Academic (oup.com)
- Considerations to address missing data when deriving clinical trial endpoints from digital health technologies
- A Markov model for inferring event types on diabetes patients data – ScienceDirect
- https://www.sciencedirect.com/science/article/pii/S0828282X20311119
- https://www.sciencedirect.com/science/article/pii/S2352914821002653
- https://towardsdatascience.com/4-techniques-to-handle-missing-values-in-time-serie s- data-c3568589b5a8

Needed skills

- Knowledge of a wide range of machine learning techniques and applications
- Experience applying machine learning algorithms and techniques, preferably to healthcare data and in particular multivariate time series. Experience with impulse trains would be desirable.
- Familiarity with real time analytics for time series prediction and dealing with missing data.
- Experience with technologies required to undertake analyses on large data sources or with computationally intensive steps (SQL, parallelization, Hadoop, Spark, HPC cluster computing, Docker) is a plus
- Fluency in statistical programming languages (R, Python, Matlab)
- Strong communication and collaboration skills
- Experience implementing reproducible research practices like version control (e.g. using Git)
- R&D mindset
- MSc/PhD degree candidate or recent graduate in Advanced Analytics related field (e.g., Statistics, Mathematics, Epidemiology, Health Economics, Outcomes Research, Computer Science, Econometrics, Physics, Engineering, Meteorology)

Mentors

Dr. Siva Chittajallu
Global Head, Algorithms and Advanced Analytics, Diabetes Care

Dr. Kamran Farooq
Senior Data Scientist @ Data & Analytics Chapter (Data Science)

Dr. Io Taxidou
Data Science Lead, Data and Analytics Chapter, Group Finance and IT

Form of cooperation

Preferred scale: 6 months full time

How to present your idea

Please formulate your ideas on how to handle missing and erroneous data, what is the state of the art and how such methods can be used to alleviate the problem in diabetes related time series. Present a short and concise slide deck (3-6 slides) to elaborate on your proposed solution. Ensure that the proposed solution is driven by state of the art/ literature review. The literature provided is indicative and you are encouraged to present methods from your own search. One page document with figures to illustrate your concept is desired in order to visualize your idea.

We don't expect a ready made solution but rather are interested in your methodology design and your approach towards solving the problem. We are looking forward to seeing your ideas and discussing your findings.