Next Generation Health Technology Assessment

Project Title: Next Generation Health Technology Assessment to support patient-centred, societally oriented, real-time decision-making on access and reimbursement for health technologies throughout Europe

Grant Agreement Number: 825162

Deliverable Title: Report of explanatory factors for treatment effectiveness based on the data from the case study. This includes experimental results using different machine learning methods showing which signals and factors are most important

Deliverable 3.2	Version: 2
Date: 12/05/2021	Lead Beneficiary: UoO
Nature: Report	Diss Level: Confidential

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

Table of Contents

DOCUMENT INFORMATION 3

DOCUMENT HISTORY 4

1. Background 5
2. Introduction 5
3. Dataset 7
4. ClinFlow – Application to visualize medical data 10
5. Understanding the structure of the data to create synthesized data 33
6. Conclusion and future work 59

Acknowledgement 60

REFERENCES 60

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

DOCUMENT INFORMATION

Grant Agreement Number	825162	Acronym	HTx
Full title	Next Generation Health Technology Assessment to support patient-centred, societally oriented, real-time decision-making on access and reimbursement for health technologies throughout Europe		
Project URL	www.htx-h2020.eu		
EU Project officer	Alexandru Padurean (Alexandru.PADUREAN@ec.europa.eu)		

Deliverable	Number	3.2	Title	Explanatory factors for treatment effectiveness
Work package	Number	3	Title	Using artificial intelligence to forecast individualised treatments on the basis of RWD

Delivery date	Contractual	31/12/2020	Actual	12/05/2021
Status		Draft Final X		
Nature		Report X Prototype o Other o		
Dissemination Level		Public X Confidential o		

Authors (Partner)	Pekka Siirtola, Juha Röning		
Responsible partner	UoO	Email	pekka.siirtola@oulu.fi

	Partner	UoO	Phone	+358 40 5181621

DOCUMENT HISTORY

NAME	DATE	VERSION	DESCRIPTION
Pekka Siirtola	08/03/2021	1.0	First draft
Pekka Siirtola	22/04/2021	1.1	Second draft
Pekka Siirtola	12/05/2021	2	Final Deliverable

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

1. Background

The data exploration and visualization phase of machine learning process concentrates on understanding the dataset and finding explanatory factors using different clustering and visualization methods. The main purpose is to get acquainted with the case study data and reveal the structures of data to select suitable AI algorithms and other methods to subsequent steps, and also to check if novel medical findings could be revealed. This deliverable is divided into two parts, and in both of these DIPP data is used. The first parts of the deliverable introduces ClinFlow, an open sourse visualization tool for medical data, which includes several methods to understand the structure of the data and visualize the data in several different ways. Therefore, ClinFlow helps to find explanatory factors from the data, and this information can be helpful in selecting the most suitable AI algorithm for the data. Moreover, to demonstrate the usefulness of ClinFlow, a data analysis study is performed using ClinFlow which shows how it can help and speed up analysis of medical data. The second part of the deliverable concentrates on studying how the understanding of the structure of the data can be used to create anonymized synthesized datasets. This topic is highly important as sharing medical datasets can be really challenging due to GDPR and privacy issues. However, if it is possible to create anonymized synthetic datasets which have structure as the original data, sharing data between different parties would be much easier.

The original idea of this deliverable was to use data from case studies but due to difficulty of sharing medical data between organizations, this deliverable was made based on DIPP data which is located at Oulu, Finland. However, ClinFlow can be used with any medical dataset and quick results can be generated with ClinFlow once the case study data will be available. In addition to the original plan, methods for generating synthetic dataset were studied, to overcome the difficulties of sharing unanonymized data in the future.

2. Introduction

AI is playing an increasingly important role in healthcare with data-driven health applications in medical diagnostics, patient monitoring, clinical decision-making, and public health policy making. AI can also be used to develop precision treatments for complex diseases and to gain insight into what makes patients healthy at the individual level. Drugs and treatments can be designed for small groups, instead of large populations, based on patients' medical history, individual variability in genes, environmental factors, lifestyle and habits, and data recorded by wearable devices.

With the availability of big data and large population-based long-term clinical follow-up studies, modern data analysis and machine learning methods have great potential of helping to uncover factors increasing the risk for the common health problems, including diabetes, heart disease, obesity and metabolic syndrome. These methods can also be used for prevention, diagnosis and prognosis of other health conditions.

The process to train AI models includes several phases: it starts with pre-processing (see deliverable 3.1.), and the next phases before actual recognition model can be trained are data exploration, statistical modelling and dimensionality reduction. The idea of these phases is to understand the structure of the data, and modify the data, for instance by creating new variables, in order to select the most appropriate methods for training the AI model. For humans, the key element in understanding the structure of the data and finding the best explanatory factors to model treatment effectiveness is visualization.

Different types of visualizations provide different information, and therefore, they can provide different types of information about the structure of the data. Moreover, when it comes to multidimensional data, like medical data, dimensionality reduction is used to avoid the course of dimensionality but also to be able to visualize the data. Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension. Working in high-dimensional spaces can be undesirable for many reasons; raw data are often sparse as a consequence of the curse of dimensionality, and analyzing the data is usually computationally intractable. [1] This means that due to dimensionality reduction, visualizations of the data are much easier to understand than visualizations without dimensionality reduction. This is important as it helps the communication between computer scientists and medical experts who have the domain knowledge. Moreover, by visualizing the data, it is easier to understand the structure of the data which helps in selecting the most suitable AI algorithms for model training. Therefore, dimensionality reduction and visualization is a crucial part of AI model training in order to achieve as good results as possible.

This deliverable focuses on understanding the structure of the data using different visualization methods. In order to provide an easy access to different visualization methods, a tool called “ClinFlow” was developed. Moreover, this deliverable shows how information of the structure understanding of the data can be used to create synthetic data sets, and why synthetic datasets

© The HTx Consortium 2019-2023. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement Nº 825162.

are important. For instance, supervised machine learning is used to compare original and synthetic data sets to show how similar they are. Finally, it explains the next steps need to be taken in order to show the potential of AI when applied to medical datasets.

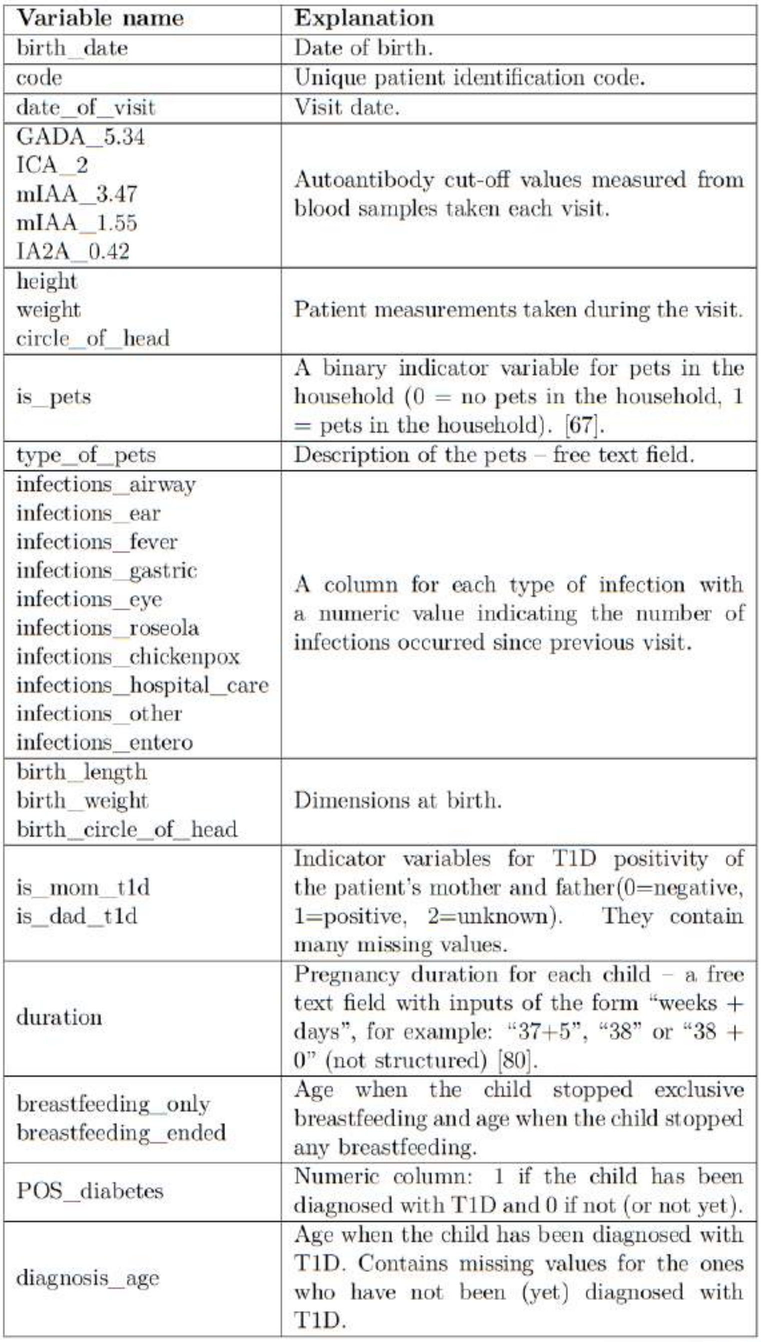
3. Dataset

Finland has the highest incidence of Type 1 Diabetes (T1D) in the world amongst young children, currently standing at approximately 72 in every 100,000 children under the age of 15 years [2]. The Type 1 Diabetes Prediction and Prevention (DIPP) Study was established in 1994 in three university hospitals in Finland to understand/learn the pathogenesis of T1D [3]. The goal of this ongoing study is to find new treatments and preventative methods by assessing risk factors in the development of T1D. The DIPP study is a population-based long-term clinical follow-up study that consists of screening newborns for increased genetic risk for diabetes.

The DIPP database used in this study has been collected since 1994 only at the Oulu University Hospital and contains information from over 6500 subjects in the form of longitudinal data; recorded since the birth of the subject. The database includes information about the subject along with the monitoring information of siblings and parents. The database also suffers from missing values due to unstandardized input methods such as information entered by hands during collection. The database comprises variables such as blood samples, infections, medications, vaccines, nutrition, and environmental factors. Blood sample data includes three autoantibody values of glutamic acid decarboxylase (GADA), protein tyrosine phosphate autoantibody (IA2A), and antibodies of insulin (IAA). Table 1 shows a list of all attributes in the raw DIPP dataset.

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

Table 1. Names and description of attributes in the raw original dataset



Variable name	Explanation
birth_date	Date of birth.
code	Unique patient identification code.
date_of_visit	Visit date.
GADA_5.34 ICA_2 mIAA_3.47 mIAA_1.55 IA2A_0.42	Autoantibody cut-off values measured from blood samples taken each visit.
height weight circle_of_head	Patient measurements taken during the visit.
is_pets	A binary indicator variable for pets in the household (0 = no pets in the household, 1 = pets in the household). [67].
type_of_pets	Description of the pets – free text field.
infections_airway infections_ear infections_fever infections_gastric infections_eye infections_roseola infections_chickenpox infections_hospital_care infections_other infections_entero	A column for each type of infection with a numeric value indicating the number of infections occurred since previous visit.
birth_length birth_weight birth_circle_of_head	Dimensions at birth.
is_mom_t1d is_dad_t1d	Indicator variables for T1D positivity of the patient's mother and father(0=negative, 1=positive, 2=unknown). They contain many missing values.
duration	Pregnancy duration for each child – a free text field with inputs of the form "weeks + days", for example: "37+5", "38" or "38 + 0" (not structured) [80].
breastfeeding_only breastfeeding_ended	Age when the child stopped exclusive breastfeeding and age when the child stopped any breastfeeding.
POS_diabetes	Numeric column: 1 if the child has been diagnosed with T1D and 0 if not (or not yet).
diagnosis_age	Age when the child has been diagnosed with T1D. Contains missing values for the ones who have not been (yet) diagnosed with T1D.

Data Cleaning

Extensive cleaning operations have been applied to the raw dataset. The following list describes in detail and motivates each cleaning action taken on the DIPP data.

1) Variables have inconsistent formats. For example, autoantibody values are stored in character form, although they are numeric values.
   a) We converted all variables to appropriate class: numeric, factor , character, date.
2) Birth date is not constant for all visits of the same participant. Some have two different birth dates (most are one day apart. One is a year apart).
   a) We replaced missing values in the birth date variable with the value that is present in other visit rows.
   b) We chose (randomly) only one birthdate for the ones that have multiple birthdates one day apart.
   c) For subjects with two different birthdays we chose the birthday closest to the first visit date.
3) Breastfeeding ending age is inconsistent, with missing values in some visit rows while present in other visit rows for the same participant.
   a) We replaced all the missing values with the value present in other visit rows for each participant.
4) Some patients have two or more different breastfeeding ending ages.
   a) We replaced all values with the maximum breastfeeding ending age present for each subject.
5) Some visit rows show that exclusive breastfeeding ended later than non exclusive breastfeeding (not possible).
   a) For these subjects, we chose the earlier value in both exclusive and non exclusive breastfeeding.
6) Mother diabetes and father diabetes columns are inconsistent and have missing values.

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

a) We converted to factor columns, with TRUE for the visit rows of patients where it's clear that the mother or father have diabetes (value = 1), and FALSE for the ones with "unknown", "0" or missing value. Here, we assumed that in the context of a T1D study, if the parents would be diagnosed with T1D, they would have definitely mentioned it. If this information is missing or unknown, it is most likely because they don't have T1D.

4. ClinFlow – Application to visualize medical data

Some key requirements for successful visualizations of large clinical datasets include dimensionality reduction, interactivity, scalability, fast results and user assistance.

ClinFlow offers a variety of methods, customized for clinical data analyses, that allow the domain expert to build cohorts and to perform exploratory analyses on large clinical datasets.

This tool is designed using the Shiny framework [40] - an R package for building interactive applications straight from R. It uses open-source technologies to offer an intuitive user interface that requires no coding or statistical software knowledge. R was chosen due to its built-in statistical analysis capabilities, and the Shiny framework makes it easy to embed R's capabilities in an interactive user interface. The application includes a module for data processing and filtering, including interactive visual elements for querying the data, checking and deleting erroneous entries, a module for exploratory analysis through a variety of graphical methods, dimensionality reduction, clustering and unsupervised learning including PCA, t-SNE, MDS, SOM, and K-means[1], a module for time-series exploratory analysis and panel data creation, and finally, a module for multivariate survival analysis using Cox proportional hazards model. Interactive visualizations were implemented using the "ggplot", "plotly" and "ggiraph" packages. These interactive plots allow users to query the data table directly by hovering, clicking or selecting points in the plots.

[1] This module is refactored from HTPdvis module of the HTPMod application by Dijun Chen [36] Source:https://github.com/htpmod/HTPmod-shinyApp

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

Our tool has a modular architecture. Removing, modifying or replacing a module in the code will not affect the rest of the functionalities. Each functionality of the application has its own server module, matched with the corresponding UI module. Figure 1 illustrates an outline of the system architecture.

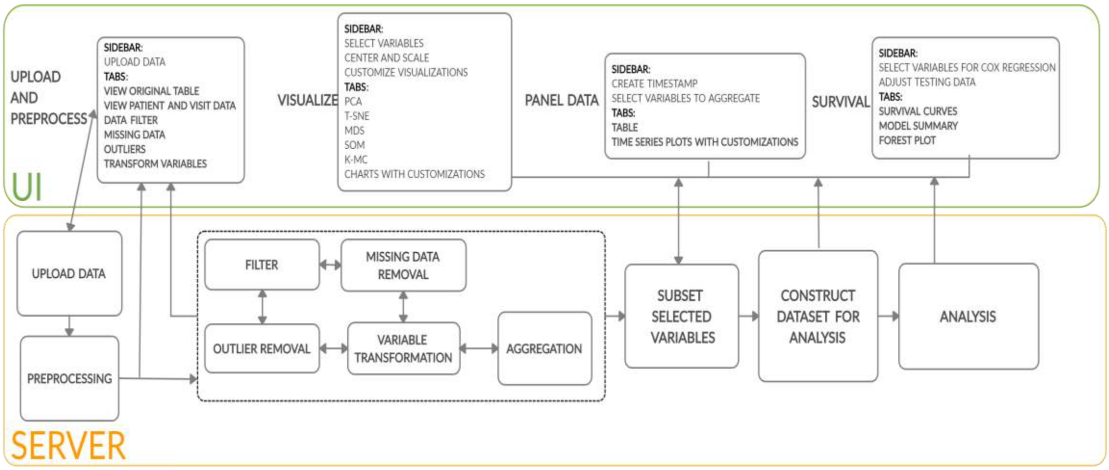
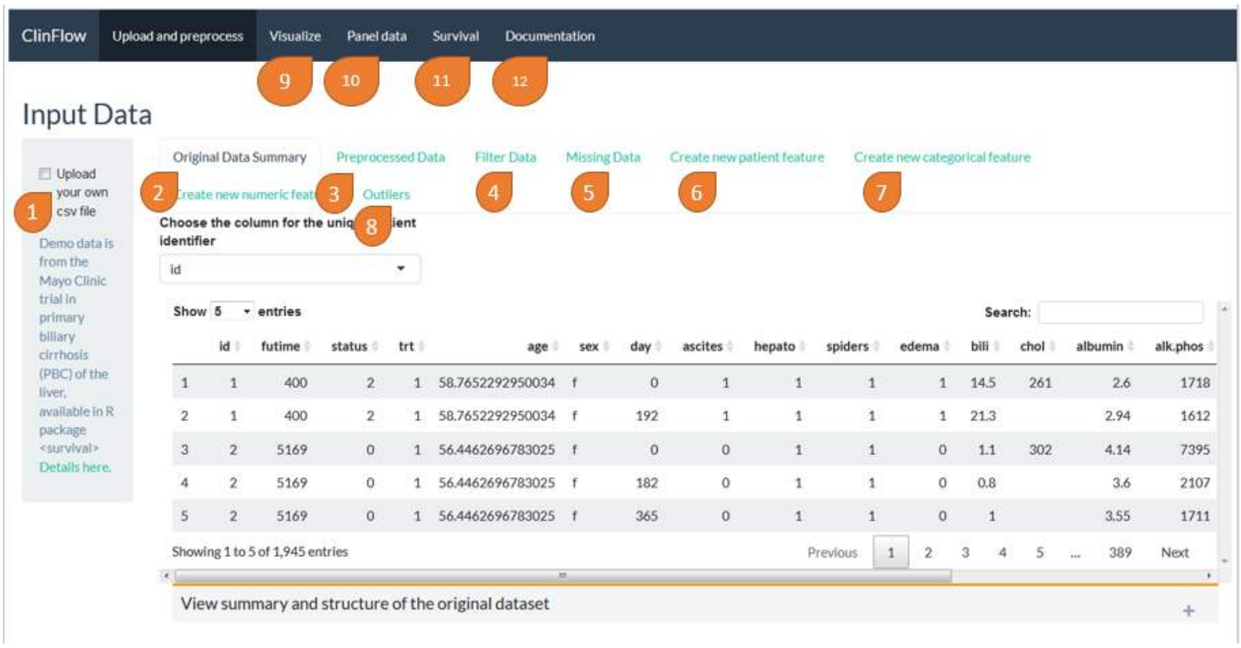

UPLOAD AND PREPROCESS
SIDEBAR: UPLOAD DATA TABS: VIEW ORIGINAL TABLE VIEW PATIENT AND VISIT DATA DATA FILTER MISSING DATA OUTLIERS TRANSFORM VARIABLES
VISUALIZE
SIDEBAR: SELECT VARIABLES CENTER AND SCALE CUSTOMIZE VISUALIZATIONS TABS: PCA T-SNE MDS SOM K-MC CHARTS WITH CUSTOMIZATIONS
PANEL DATA
SIDEBAR: CREATE TIMESTAMP SELECT VARIABLES TO AGGREGATE TABS: TABLE TIME SERIES PLOTS WITH CUSTOMIZATIONS
SURVIVAL
SIDEBAR: SELECT VARIABLES FOR COX REGRESSION ADJUST TESTING DATA TABS: SURVIVAL CURVES MODEL SUMMARY FOREST PLOT
UI
UPLOAD DATA
PREPROCESSING
FILTER
MISSING DATA REMOVAL
OUTLIER REMOVAL
VARIABLE TRANSFORMATION
AGGREGATION
SUBSET SELECTED VARIABLES
CONSTRUCT DATASET FOR ANALYSIS
ANALYSIS
SERVER

Figure 1. Architecture of ClinFlow Shiny app.

An online demo of the ClinFlow can be found here. In this demo we used data from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver, available in R package “survival” Details here. Source code of this application can be found here. Therefore, as ClinFlow is an open source application, user can also add more data analysis and visualization methods to is him/herself if it is missing some feature that is important to the user.

4.1. Using ClinFlow

Figure 2 shows the main UI of the ClinFlow, and this section explains what functionalities it has, and how these can be used.



Input Data

☐ Upload your own csv file

Demo data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver, available in R package <survival>. Details here.

Choose the column for the uniq ient identifier: id

	id	futime	status	trt	age	sex	day	ascites	hepato	spiders	edema	bili	chol	albumin	alk.phos
1	1	400	2	1	58.7652292950034	f	0	1	1	1	1	14.5	261	2.6	1718
2	1	400	2	1	58.7652292950034	f	192	1	1	1	1	21.3		2.94	1612
3	2	5169	0	1	56.4462696783025	f	0	0	1	1	0	1.1	302	4.14	7395
4	2	5169	0	1	56.4462696783025	f	182	0	1	1	0	0.8		3.6	2107
5	2	5169	0	1	56.4462696783025	f	365	0	1	1	0	1		3.55	1711

Showing 1 to 5 of 1,945 entries

View summary and structure of the original dataset

Figure 2. ClinFlow main page.

1. Input

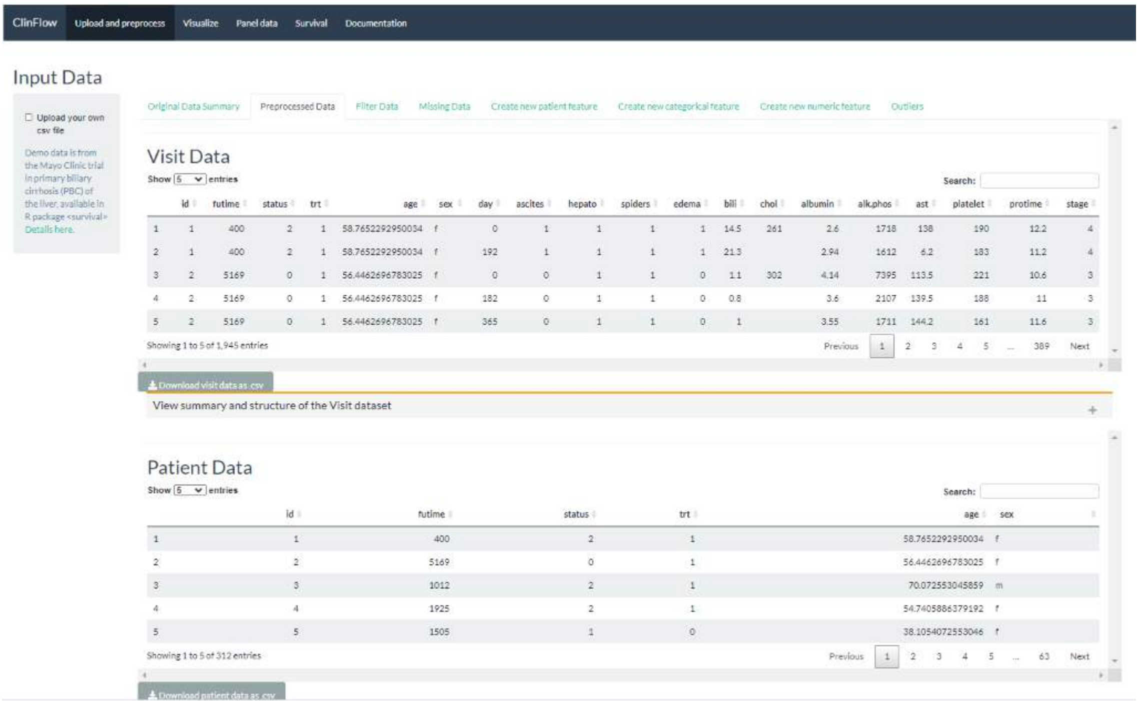
The application expects a .csv or .tsv file as input.

2. Original data

In the Original Data Summary tab, a sample of the first rows of the uploaded raw table is displayed, along with a box that displays a summary and structure of the data.

3. Preprocessed data

The following tab, Preprocessed Data displays two tables, Visit data and Patient data, along with buttons to download or display the summary and structure of each table. Patient data contains only the constant (nontime series) variables, such as patient's sex. Visit data contains all variables (time-series and constant), as illustrated in Figure 3.



Input Data

☐ Upload your own csv file

Demo data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver, available in R package <survival>. Details here.

Original Data Summary
Preprocessed Data
Filter Data
Missing Data
Create new patient feature
Create new categorical feature
Create new numeric feature
Outliers

Visit Data

Show 5 entries

Search:

	id	futime	status	trt	age	sex	day	ascites	hepato	spiders	edema	bili	chol	albumin	alk.phos	ast	platelet	protime	stage
1	1	400	2	1	58.7652292950034	f	0	1	1	1	1	14.5	261	2.6	1718	138	190	12.2	4
2	1	400	2	1	58.7652292950034	f	192	1	1	1	1	21.3		2.94	1612	6.2	183	11.2	4
3	2	5169	0	1	56.4462696783025	f	0	0	1	1	0	1.1	302	4.14	7395	113.5	221	10.6	3
4	2	5169	0	1	56.4462696783025	f	182	0	1	1	0	0.8		3.6	2107	139.5	188	11	3
5	2	5169	0	1	56.4462696783025	f	365	0	1	1	0	1		3.55	1711	144.2	161	11.6	3

Showing 1 to 5 of 1,945 entries

View summary and structure of the Visit dataset

Patient Data

Show 5 entries

Search:

	id	futime	status	trt	age	sex
1	1	400	2	1	58.7652292950034	f
2	2	5169	0	1	56.4462696783025	f
3	3	1012	2	1	70.072553045859	m
4	4	1925	2	1	54.7405886379192	f
5	5	1505	1	0	38.1054072553046	f

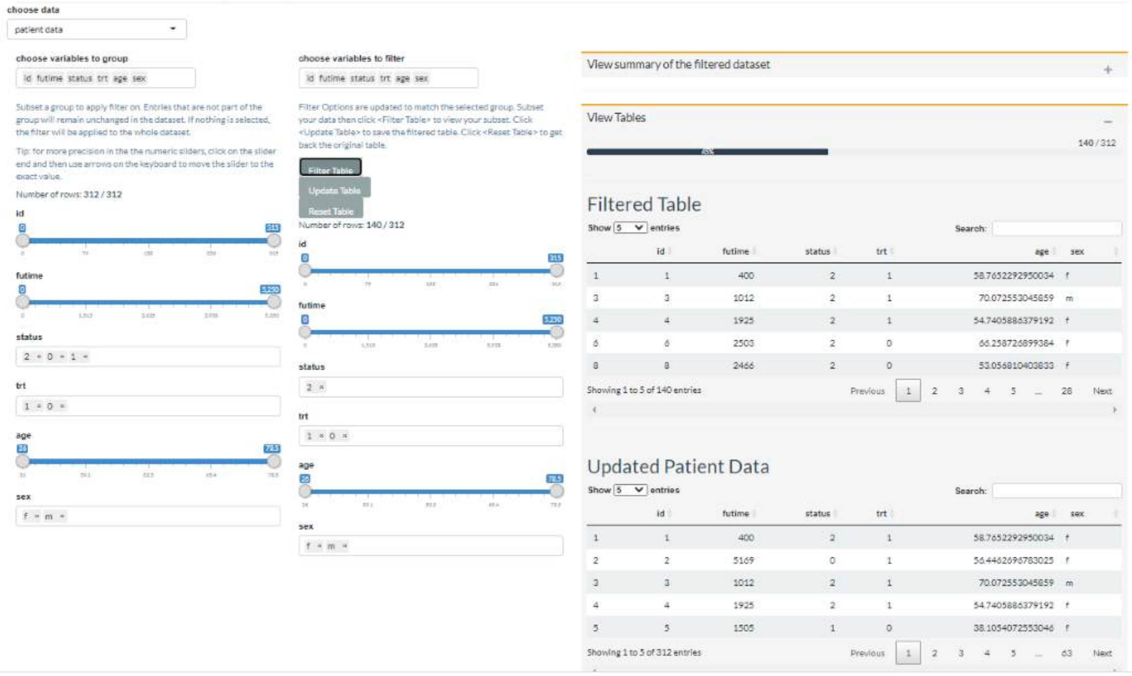
Showing 1 to 5 of 312 entries

Figure 3. Preprocessed data tables.

4. Filter

The next tab is the Filter Data tab, a functionality that allows the user to choose either visit data or patient data, and to select variables to group and filter. The filter elements on the left are used to group the data, a pre-filter. This is used to create a subset of data to apply the filter on, while the rest of the data remains unchanged. For the numeric variables, the filtering is done via a slider, and for the categorical ones, a multi-choice selector. If the variables contain missing entries, there is also a switch for filtering out the rows with missing values in the chosen variables. The filtered table is updated dynamically as the user filters the data, and the bar above the table shows in percents the amount of data preserved after filtering. The button updates both the visit and patient data displayed below, to include only the filtered information and it also updates the filter options accordingly. There is an option for resetting the table, which brings back the unfiltered preprocessed table from the start, and resets the filter options. This page also contains a box for displaying a summary of the variables and the variable types of the filtered data, and buttons for downloading the updated patient and visit tables. The filter is illustrated in Figure 4.

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.



choose data: patient data

choose variables to group: id futime status trt age sex

Subset a group to apply filter on. Entries that are not part of the group will remain unchanged in the dataset. If nothing is selected, the filter will be applied to the whole dataset.

Tip: for more precision in the the numeric sliders, click on the slider end and then use arrows on the keyboard to move the slider to the exact value.

Number of rows: 312 / 312

id

futime

status: 2 × 0 × 1 ×

trt: 1 × 0 ×

age

sex: f × m ×

choose variables to filter: id futime status trt age sex

Filter Options are updated to match the selected group. Subset your data then click <Filter Table> to view your subset. Click <Update Table> to save the filtered table. Click <Reset Table> to get back the original table.

Number of rows: 140 / 312

id

futime

status: 2 ×

trt: 1 × 0 ×

age

sex: f × m ×

View summary of the filtered dataset

View Tables

140 / 312

Filtered Table

Show 5 entries Search:

	id	futime	status	trt	age	sex
1	1	400	2	1	58.7652292950034	f
3	3	1012	2	1	70.072553045859	m
4	4	1925	2	1	54.7405886379192	f
6	6	2503	2	0	66.258726899384	f
8	8	2466	2	0	53.056810403833	f

Showing 1 to 5 of 140 entries Previous 1 2 3 4 5 … 28 Next

Updated Patient Data

Show 5 entries Search:

	id	futime	status	trt	age	sex
1	1	400	2	1	58.7652292950034	f
2	2	5169	0	1	56.4462696783025	f
3	3	1012	2	1	70.072553045859	m
4	4	1925	2	1	54.7405886379192	f
5	5	1505	1	0	38.1054072553046	f

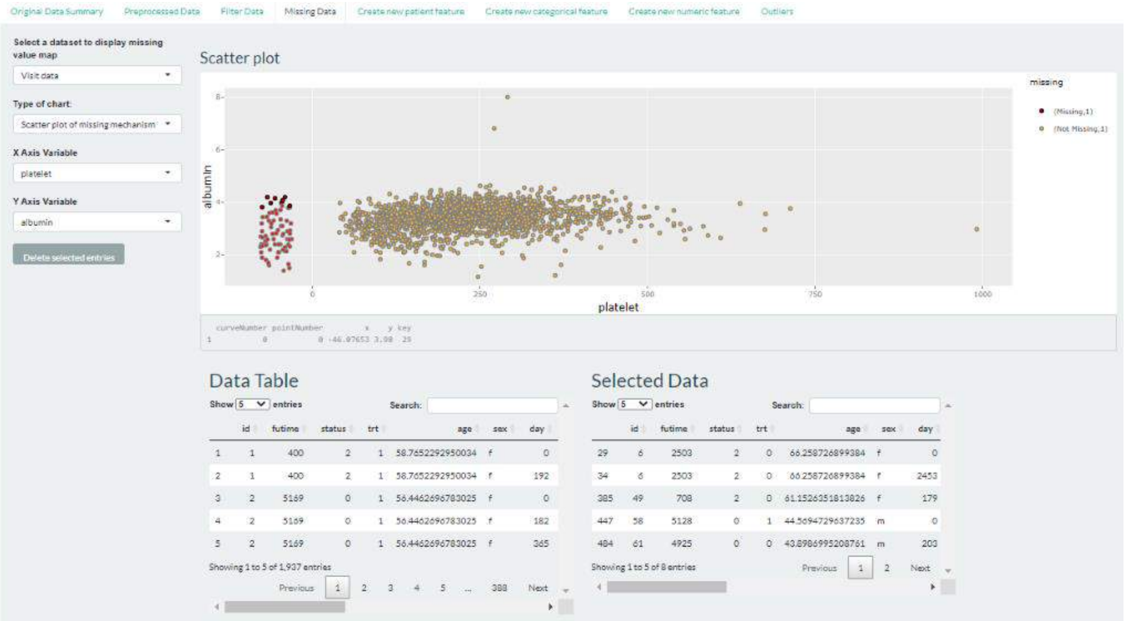
Showing 1 to 5 of 312 entries Previous 1 2 3 4 5 … 63 Next

Figure 4. Data filter.

5. Missing data

In this tab, the user can choose to display a missing data map of either the visit or the patient table, and a scatterplot of missing vs. observed values from two chosen variables, in order to study the missing data mechanism. This plot can show whether the data is missing completely at random (MCAR), or not. If the data is not MCAR, this table helps the user to study the missing mechanism and plan on how to deal with the missing values without introducing bias in the analysis results. The interactive plot allows to select the points by clicking or dragging, then display them in a separate table, and delete them via the button (Figure 5).

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.



Select a dataset to display missing value map: Visit data

Type of chart: Scatter plot of missing mechanism

X Axis Variable: platelet

Y Axis Variable: albumin

```
  curveNumber pointNumber         x    y key
1           0           0 -46.07653 3.08  25
```

Data Table

Show 5 entries Search:

	id	futime	status	trt	age	sex	day
1	1	400	2	1	58.7652292950034	f	0
2	1	400	2	1	58.7652292950034	f	192
3	2	5169	0	1	56.4462696783025	f	0
4	2	5169	0	1	56.4462696783025	f	182
5	2	5169	0	1	56.4462696783025	f	365

Showing 1 to 5 of 1,937 entries

Previous 1 2 3 4 5 ... 388 Next

Selected Data

Show 5 entries Search:

	id	futime	status	trt	age	sex	day
29	6	2503	2	0	66.258726899384	f	0
34	6	2503	2	0	66.258726899384	f	2453
385	49	708	2	0	61.1526351813826	f	179
447	58	5128	0	1	44.5694729637235	m	0
484	61	4925	0	0	43.8986995208761	m	203

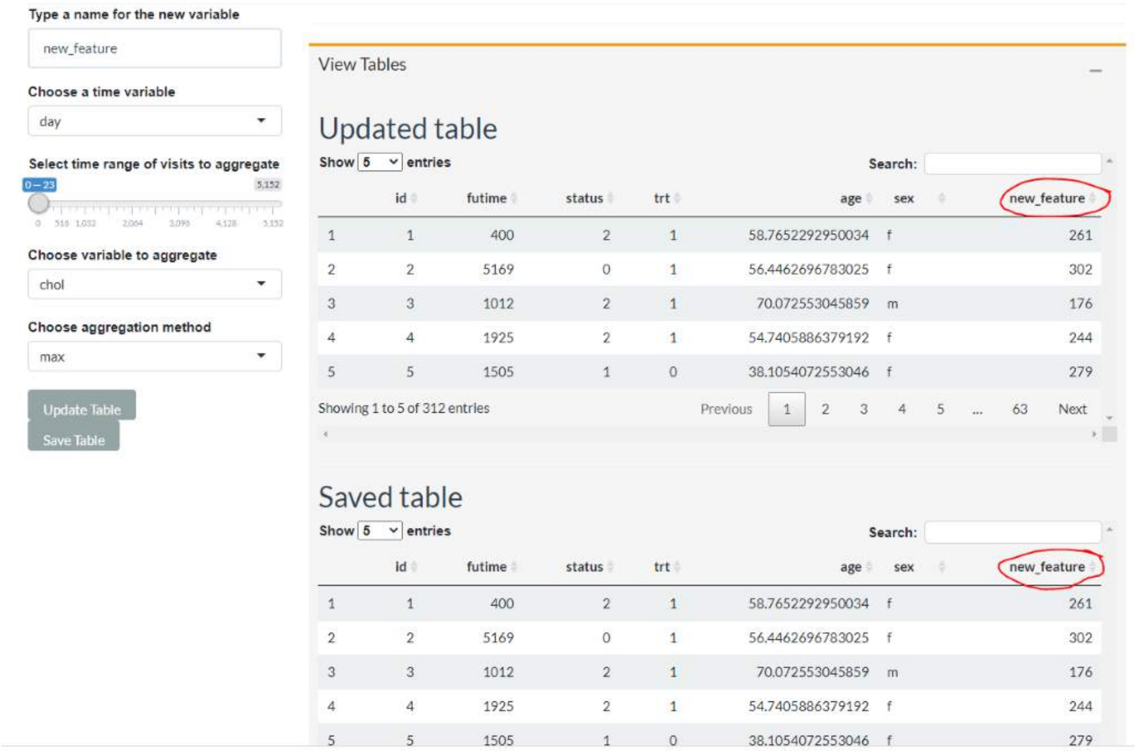
Showing 1 to 5 of 8 entries

Previous 1 2 Next

Figure 5. Missing data interactive plot with a few points selected(outlined with black)

6. Create a new patient feature

This option allows the user to aggregate time-series data from a certain period of time, and attach it to the Patient table as a constant variable. The options for aggregation of numeric variables are: sum, max, min, mean (Figure 6). For the categorical variables, user is asked to select a categorical level. If that level appears in the time interval, the variable will have the value TRUE, otherwise FALSE. For example: If the patient had a respiratory infection during days 1-10, mark TRUE, otherwise FALSE.



Type a name for the new variable: new_feature

Choose a time variable: day

Select time range of visits to aggregate: 0 – 23

Choose variable to aggregate: chol

Choose aggregation method: max

Updated table

Show 5 entries Search:

	id	futime	status	trt	age	sex	new_feature
1	1	400	2	1	58.7652292950034	f	261
2	2	5169	0	1	56.4462696783025	f	302
3	3	1012	2	1	70.072553045859	m	176
4	4	1925	2	1	54.7405886379192	f	244
5	5	1505	1	0	38.1054072553046	f	279

Showing 1 to 5 of 312 entries

Saved table

Show 5 entries Search:

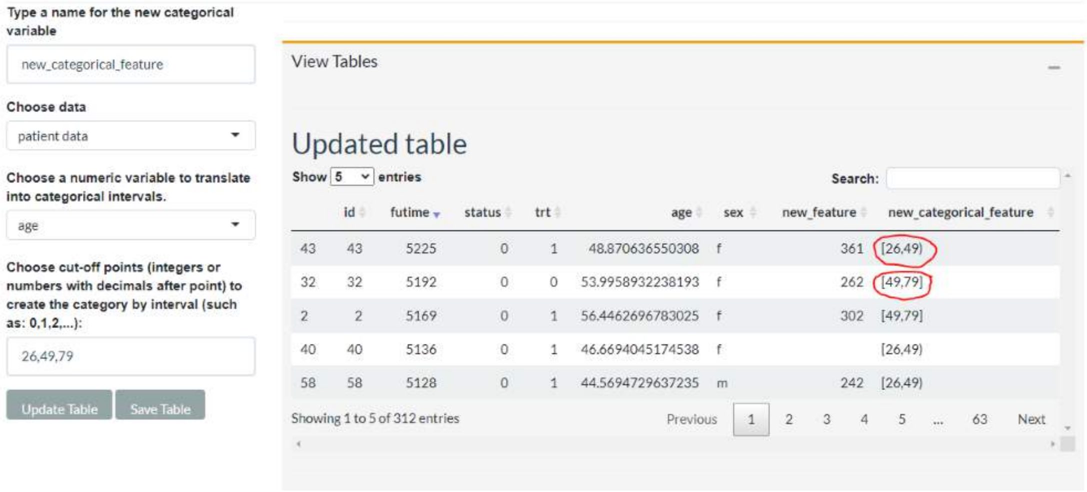
	id	futime	status	trt	age	sex	new_feature
1	1	400	2	1	58.7652292950034	f	261
2	2	5169	0	1	56.4462696783025	f	302
3	3	1012	2	1	70.072553045859	m	176
4	4	1925	2	1	54.7405886379192	f	244
5	5	1505	1	0	38.1054072553046	f	279

Figure 6. Created a new patient feature by aggregating visit cholesterol data for the visit days 0-43.

7. Create a new categorical feature

This tab allows the user to create a new feature, either in the patient table or the visit table, based on another numerical feature in that table. The user must choose a numerical variable in the drop down menu, then type the cut-off points, separated by comma, for splitting the variable into intervals closed on the left and open on the right. The last interval is closed on both sides. A new categorical variable is created with as many levels as there are intervals. The displays a preview of the table containing the new categorical variable, and the button adds the new variable to the dataset. The new feature can then be used for conducting group-based analysis(Figure 7).

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.



Type a name for the new categorical variable

new_categorical_feature

Choose data

patient data

Choose a numeric variable to translate into categorical intervals.

age

Choose cut-off points (integers or numbers with decimals after point) to create the category by interval (such as: 0,1,2,...):

26,49,79

Updated table

Show 5 entries Search:

	id	futime	status	trt	age	sex	new_feature	new_categorical_feature
43	43	5225	0	1	48.870636550308	f	361	[26,49)
32	32	5192	0	0	53.9958932238193	f	262	[49,79]
2	2	5169	0	1	56.4462696783025	f	302	[49,79]
40	40	5136	0	1	46.6694045174538	f		[26,49)
58	58	5128	0	1	44.5694729637235	m	242	[26,49)

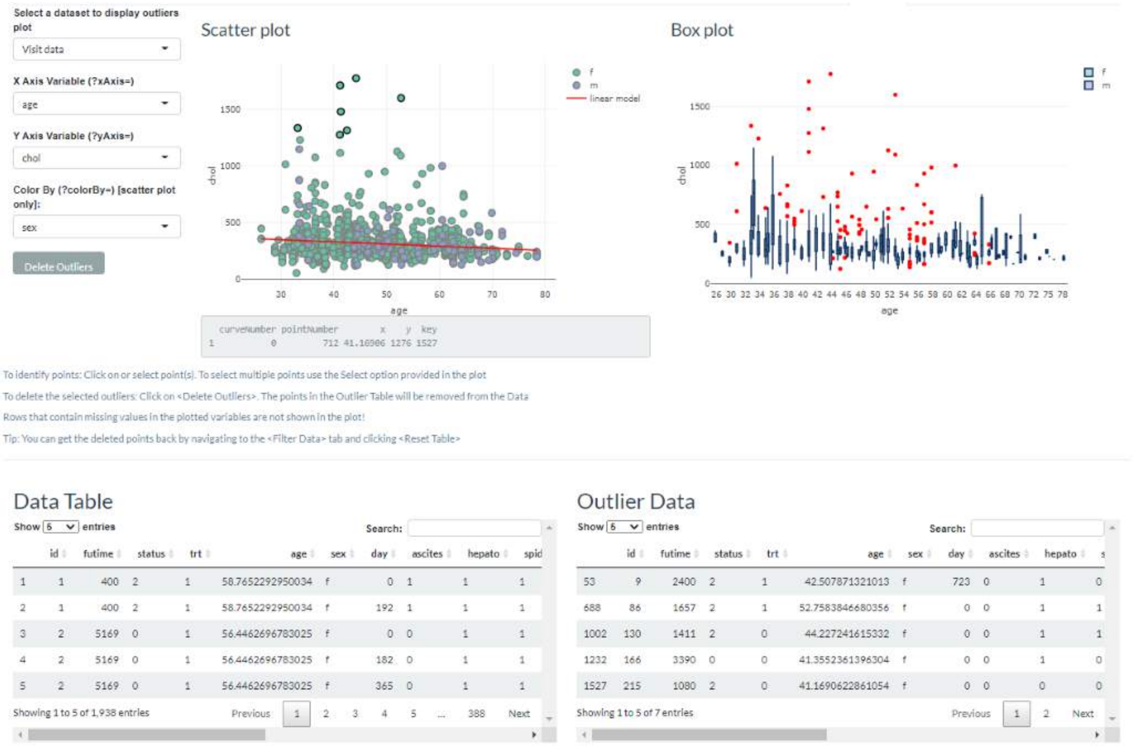
Showing 1 to 5 of 312 entries

Figure 7. Created a new categorical feature by splitting the age variable into intervals.

8. Outliers

This tab allows the user to plot a scatterplot and a boxplot of two chosen variables, and color the points by a categorical variable, in order to identify outliers. The scatterplot shows the points in the data and a regression line. The points that are further from the regression line should be investigated as possible outliers. The boxplot shows a representation of the distribution of values on the Y axis, as a box with the edges as the first and third quartile and a median line in between. If the X axis is a categorical variable, the boxplot shows distributions of the data for each category. If the X axis is continuous, the boxplot creates boxes for each value (Figure 8). The values that are far away from the median and the quartiles should be investigated as outliers. The data can be either patient or visit data, and the plots are interactive. Clicking on a point in the scatterplot, or selecting multiple points by dragging, will move those entries from the original table into the Outlier table, where the user can study them and decide whether they should be removed from the data or not. Once the user presses the button, the new data table without outliers is saved, and the selected entries are deleted (Figure 8).

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.



```
  curveNumber pointNumber        x    y  key
1           0         712 41.16906 1276 1527
```

To identify points: Click on or select point(s). To select multiple points use the Select option provided in the plot

To delete the selected outliers: Click on <Delete Outliers>. The points in the Outlier Table will be removed from the Data

Rows that contain missing values in the plotted variables are not shown in the plot!

Tip: You can get the deleted points back by navigating to the <Filter Data> tab and clicking <Reset Table>

Data Table

Show 5 entries Search:

	id	futime	status	trt	age	sex	day	ascites	hepato	spid
1	1	400	2	1	58.7652292950034	f	0	1	1	1
2	1	400	2	1	58.7652292950034	f	192	1	1	1
3	2	5169	0	1	56.4462696783025	f	0	0	1	1
4	2	5169	0	1	56.4462696783025	f	182	0	1	1
5	2	5169	0	1	56.4462696783025	f	365	0	1	1

Showing 1 to 5 of 1,938 entries Previous 1 2 3 4 5 … 388 Next

Outlier Data

Show 5 entries Search:

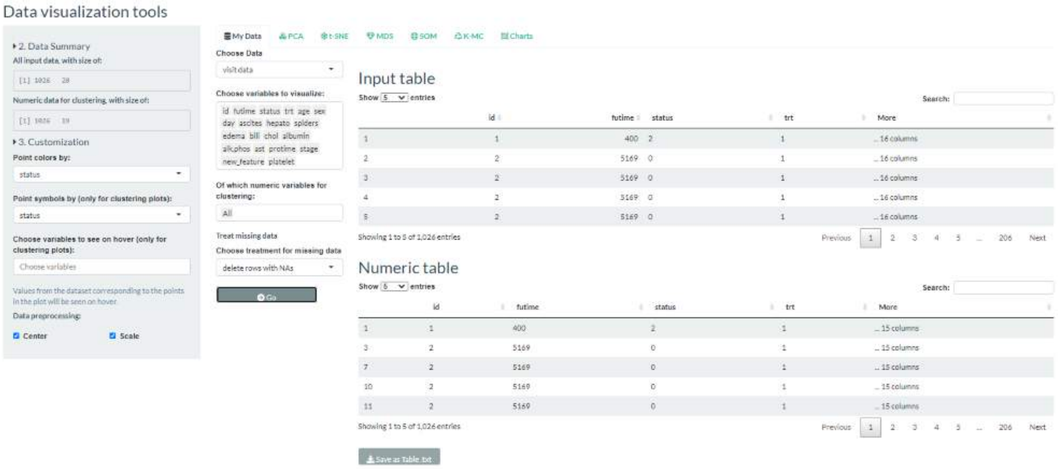
	id	futime	status	trt	age	sex	day	ascites	hepato	s
53	9	2400	2	1	42.507871321013	f	723	0	1	0
688	86	1657	2	1	52.7583846680356	f	0	0	1	1
1002	130	1411	2	0	44.227241615332	f	0	0	1	1
1232	166	3390	0	0	41.3552361396304	f	0	0	1	0
1527	215	1080	2	0	41.1690622861054	f	0	0	0	0

Showing 1 to 5 of 7 entries Previous 1 2 Next

Figure 8. Outliers interactive plot with a few points selected (outlined with black)

9. Visualize

The second navigation bar includes a tab for constructing and viewing the table to be analyzed, tabs for different clustering methods, and a tab for various charts. In the My data tab (Figure 9), the user can select either the patient or the visit table, then choose the variables and missing data treatment to use in the clustering methods. The numeric variables are used for the clustering, and the categorical variables can be used to customize the colors and the shapes of the points in the cluster plots. The user can choose how to treat missing values in the numeric variables used for clustering, either by deleting the rows containing missing values, estimating the missing values using Bayesian PCA, or not treating them at all. The clustering methods give an error if missing values are present in the numeric data, so the user must make an informed decision on how to proceed. It is recommended to study the missing mechanisms in the data and make a subset, using the data filter, that doesn't include missing values that are not MCAR. Imputation or deletion of the missing entries that are not MCAR can introduce bias in the analysis. If the categorical variables contain missing values, the NA entries are automatically assigned the label and they appear as a category in the data, in order to preserve as much information as possible. The clustering visualizations(Figures 10-13) allow for user customization of some parameters, rotation of the 3D plot, and saving the plots in various formats. Clustering is useful in identifying groups of similar entries in the table. Combined with the coloring and categorization options, the user can explore the reasons behind the similarities found in the data points and identify important relationships between variables.



Data visualization tools

2. Data Summary

All input data, with size of: [1] 1026 28

Numeric data for clustering, with size of: [1] 1026 19

3. Customization

Point colors by: status

Point symbols by (only for clustering plots): status

Choose variables to see on hover (only for clustering plots): Choose variables

Values from the dataset corresponding to the points in the plot will be seen on hover

Data preprocessing:

☑ Center ☑ Scale

Choose Data: visit data

Choose variables to visualize: id futime status trt age sex day ascites hepato spiders edema bili chol albumin alk.phos ast protime stage new_feature platelet

Of which numeric variables for clustering: All

Treat missing data

Choose treatment for missing data: delete rows with NAs

Input table

Show 5 entries Search:

	id	futime	status	trt	More
1	1	400	2	1	... 16 columns
2	2	5169	0	1	... 16 columns
3	2	5169	0	1	... 16 columns
4	2	5169	0	1	... 16 columns
5	2	5169	0	1	... 16 columns

Showing 1 to 5 of 1,026 entries Previous 1 2 3 4 5 ... 206 Next

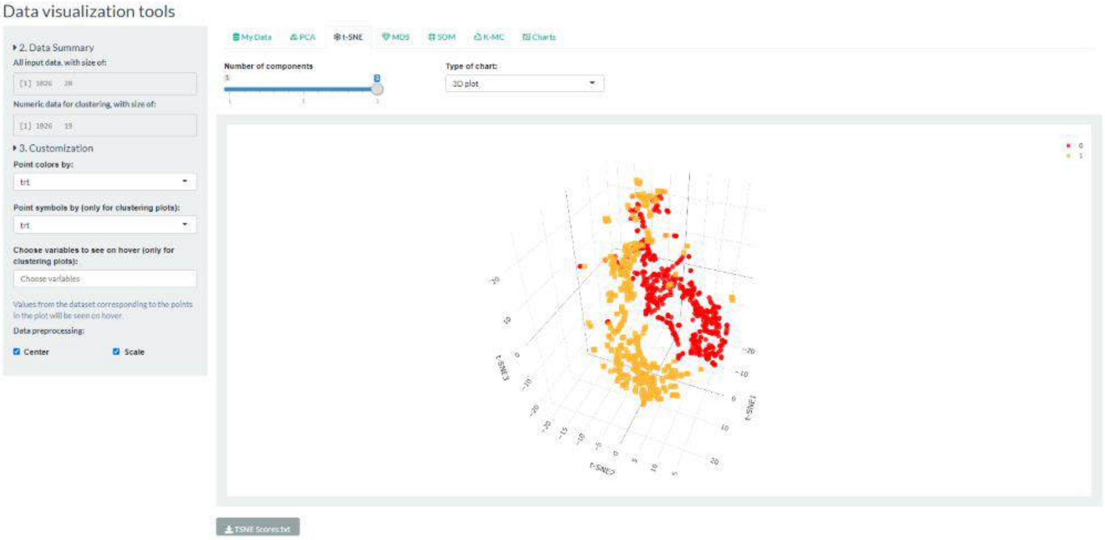
Numeric table

Show 5 entries Search:

	id	futime	status	trt	More
1	1	400	2	1	... 15 columns
3	2	5169	0	1	... 15 columns
7	2	5169	0	1	... 15 columns
10	2	5169	0	1	... 15 columns
11	2	5169	0	1	... 15 columns

Showing 1 to 5 of 1,026 entries Previous 1 2 3 4 5 ... 206 Next

Figure 9. The dataset constructed for analysis



Data visualization tools

2. Data Summary

All input data, with size of: [1] 1026 28

Numeric data for clustering, with size of: [1] 1026 19

3. Customization

Point colors by: trt

Point symbols by (only for clustering plots): trt

Choose variables to see on hover (only for clustering plots): Choose variables

Values from the dataset corresponding to the points in the plot will be seen on hover.

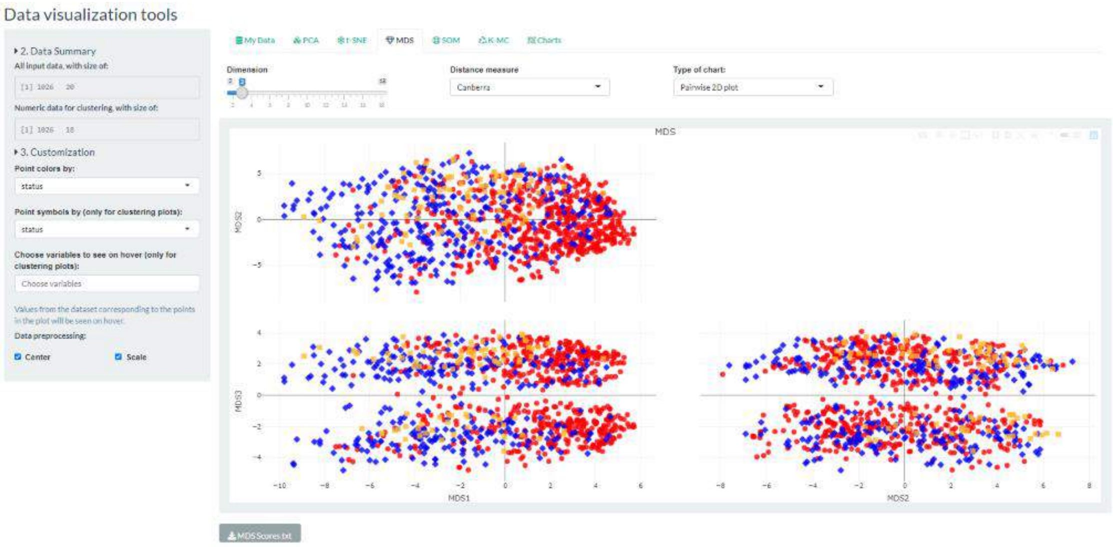
Data preprocessing:

☑ Center ☑ Scale

Number of components: 3

Type of chart: 3D plot

Figure 10. 3D T-sne plot



Data visualization tools

2. Data Summary

All input data, with size of:

```
[1] 1026   20
```

Numeric data for clustering, with size of:

```
[1] 1026   18
```

3. Customization

Point colors by: status

Point symbols by (only for clustering plots): status

Choose variables to see on hover (only for clustering plots): Choose variables

Values from the dataset corresponding to the points in the plot will be seen on hover.

Data preprocessing:

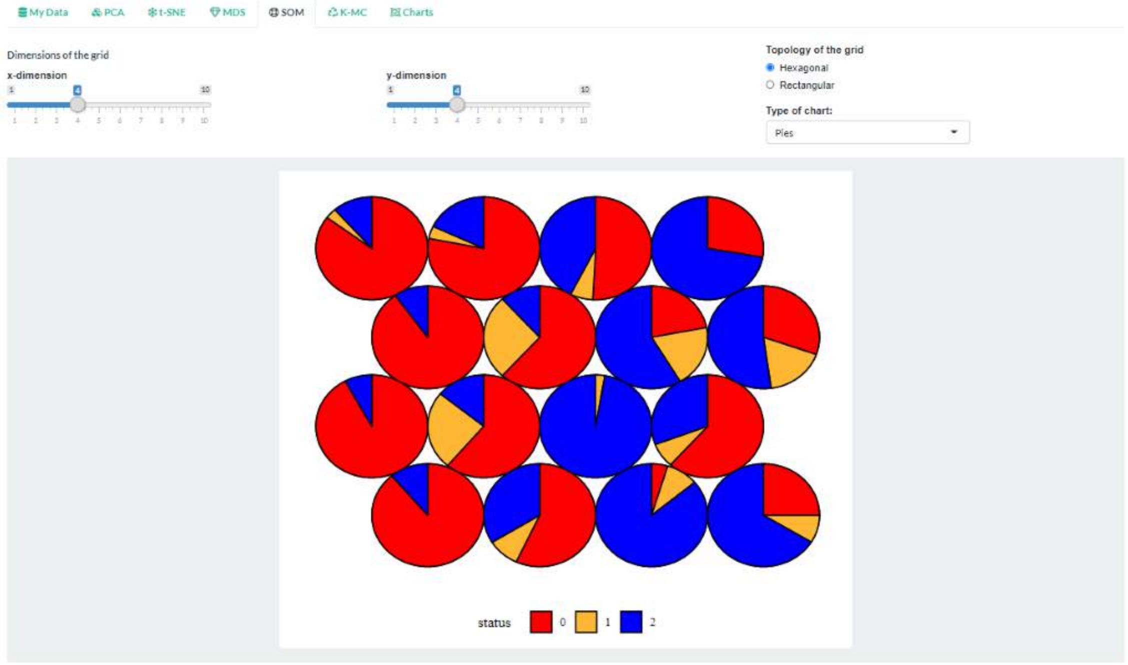
☑ Center ☑ Scale

Dimension: 3

Distance measure: Canberra

Type of chart: Pairwise 2D plot

Figure 11. Pairwise 2D MDS plot.



Dimensions of the grid

x-dimension: 4

y-dimension: 4

Topology of the grid

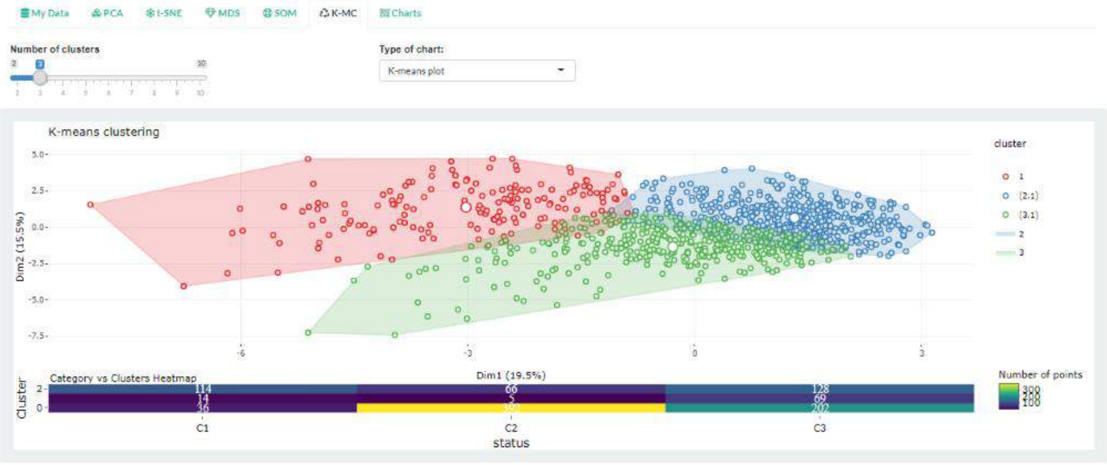
● Hexagonal

○ Rectangular

Type of chart: Pies

Figure 12. Self-organizing map pies plot.

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.



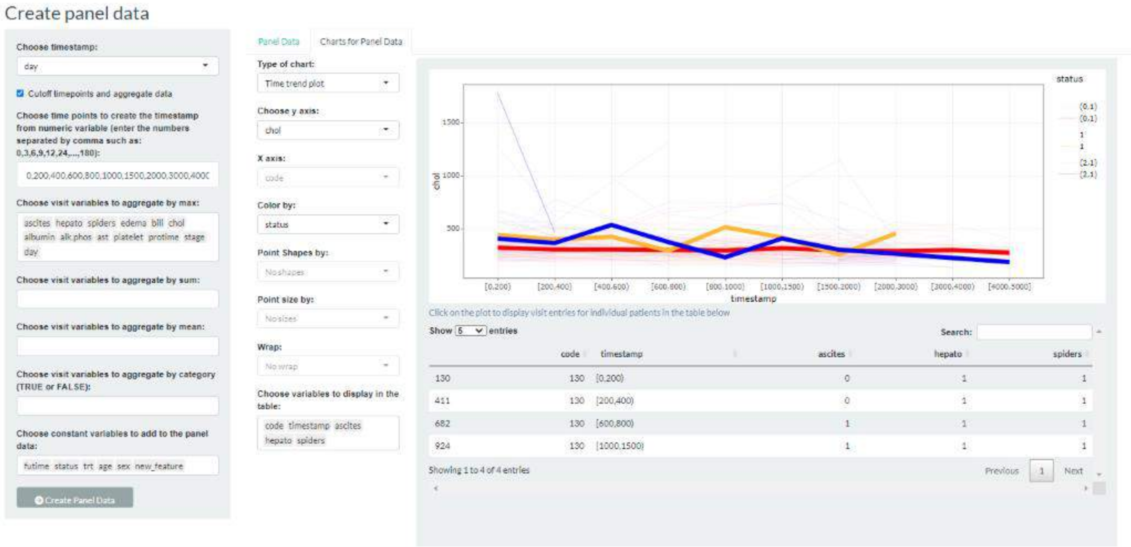
Number of clusters: 3

Type of chart: K-means plot

Figure 13. K-Means clustering plot.

10. Panel Data

This is a panel data creation tool that turns the visit time into a timestamp and allows the user to construct a time-series longitudinal panel according to preferences. User can either cut off time points and aggregate data for each patient at each time interval, or keep all the time points. The Panel data charts offer interactive visualizations for time-series values(Figures 14-15). Clicking on the points or lines in the plots, representing the values of individual patients, will display a table with all the entries of that patient.



Create panel data

Choose timestamp: day

☑ Cutoff timepoints and aggregate data

Choose time points to create the timestamp from numeric variable (enter the numbers separated by comma such as: 0,3,6,9,12,24,...,180): 0,200,400,600,800,1000,1500,2000,3000,400(

Choose visit variables to aggregate by max: ascites hepato spiders edema bili chol albumin alk.phos ast platelet protime stage day

Choose visit variables to aggregate by sum:

Choose visit variables to aggregate by mean:

Choose visit variables to aggregate by category (TRUE or FALSE):

Choose constant variables to add to the panel data: futime status trt age sex new_feature

Type of chart: Time trend plot

Choose y axis: chol

X axis: code

Color by: status

Point Shapes by: No shapes

Point size by: No sizes

Wrap: No wrap

Choose variables to display in the table: code timestamp ascites hepato spiders

Click on the plot to display visit entries for individual patients in the table below

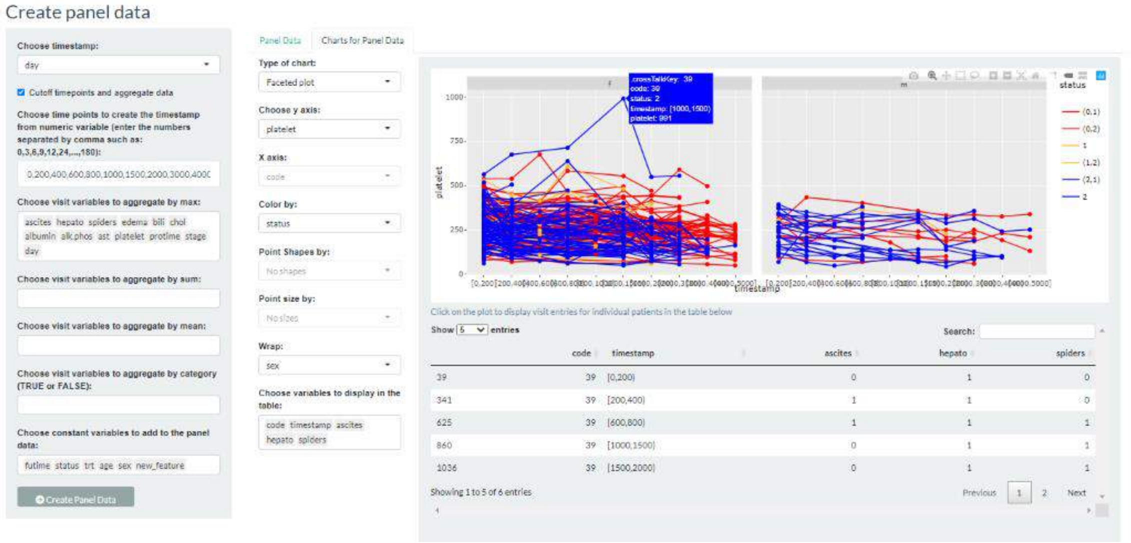
Show 5 entries

Search:

	code	timestamp	ascites	hepato	spiders
130	130	[0,200)	0	1	1
411	130	[200,400)	0	1	1
682	130	[600,800)	1	1	1
924	130	[1000,1500)	1	1	1

Showing 1 to 4 of 4 entries

Figure 14. Time series plot for cholesterol values colored by status.



Create panel data

Choose timestamp: day

☑ Cutoff timepoints and aggregate data

Choose time points to create the timestamp from numeric variable (enter the numbers separated by comma such as: 0,3,6,9,12,24,...,180): 0,200,400,600,800,1000,1500,2000,3000,400(

Choose visit variables to aggregate by max: ascites hepato spiders edema bili chol albumin alk.phos ast platelet protime stage day

Choose visit variables to aggregate by sum:

Choose visit variables to aggregate by mean:

Choose visit variables to aggregate by category (TRUE or FALSE):

Choose constant variables to add to the panel data: futime status trt age sex new_feature

Type of chart: Faceted plot

Choose y axis: platelet

X axis: code

Color by: status

Point Shapes by: No shapes

Point size by: No sizes

Wrap: sex

Choose variables to display in the table: code timestamp ascites hepato spiders

Click on the plot to display visit entries for individual patients in the table below

Show 5 entries

Search:

	code	timestamp	ascites	hepato	spiders
39	39	[0,200)	0	1	0
341	39	[200,400)	1	1	0
625	39	[600,800)	1	1	1
860	39	[1000,1500)	0	1	1
1036	39	[1500,2000)	0	1	1

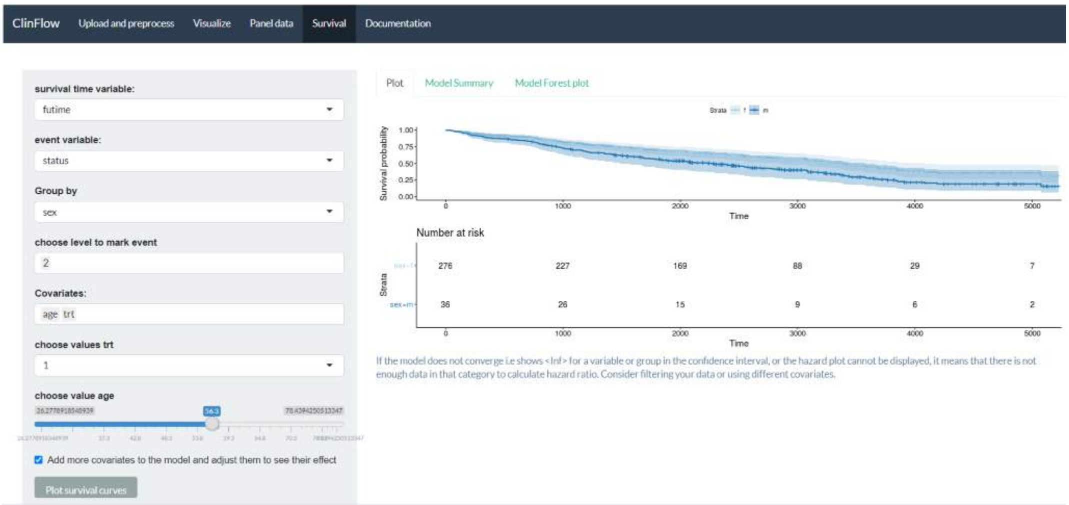
Showing 1 to 5 of 6 entries

Figure 15. Time series plot for platelet values colored by status, wrapped by sex.

11. Survival

This functionality allows the user to plot survival curves with a Cox PH model and explore the effect of different covariates on the survival. In the left panel, the user selects the Time-to-event variable (or follow-up time), an event variable and event value, and more covariates if needed.

The app fits a coxPH model on the data, then predicts the survival times of a new patient with the chosen covariates values. The user can explore the effect of different covariates on the survival curves. The model details can be investigated in the Model Summary tab, and a Hazard forest plot can be visualized in the Model Forest Plot tab(Figures 16-18).



survival time variable: futime

event variable: status

Group by sex

choose level to mark event 2

Covariates: age trt

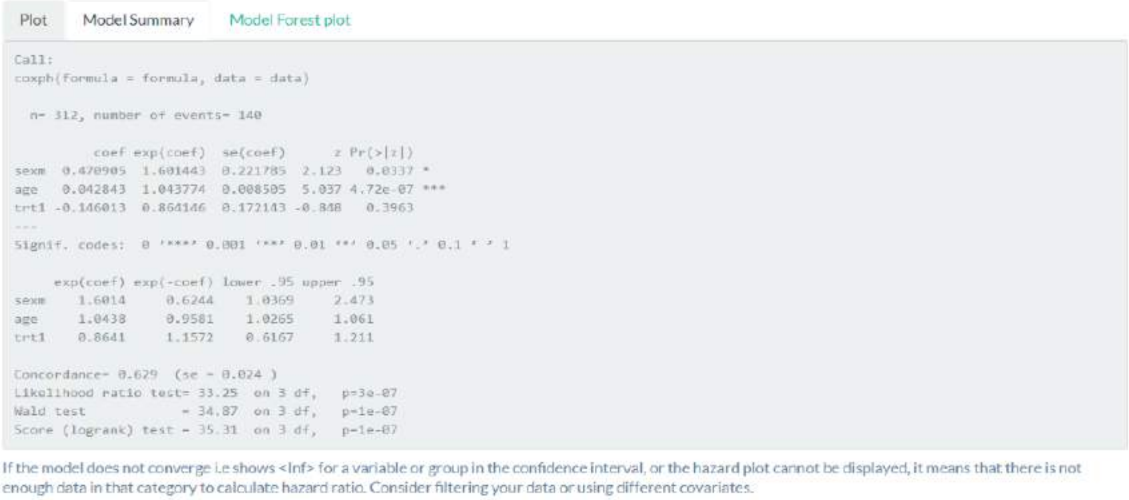
choose values trt 1

choose value age 26.2778918548939 56.3 78.4394250513347

☑ Add more covariates to the model and adjust them to see their effect

If the model does not converge i.e shows <Inf> for a variable or group in the confidence interval, or the hazard plot cannot be displayed, it means that there is not enough data in that category to calculate hazard ratio. Consider filtering your data or using different covariates.

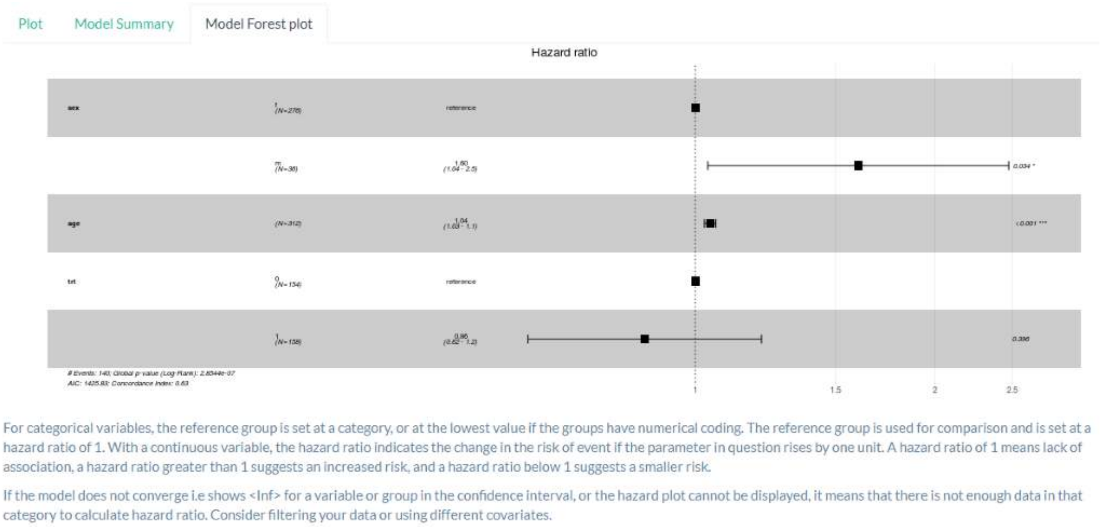
Figure 16. Survival time for males vs females of 56 years old who received treatment.



```
Call:
coxph(formula = formula, data = data)

  n= 312, number of events= 140

         coef exp(coef)  se(coef)      z Pr(>|z|)
sexm  0.470905  1.601443  0.221785  2.123   0.0337 *
age   0.042843  1.043774  0.008505  5.037 4.72e-07 ***
trt1 -0.146013  0.864146  0.172143 -0.848   0.3963
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     exp(coef) exp(-coef) lower .95 upper .95
sexm    1.6014     0.6244    1.0369     2.473
age     1.0438     0.9581    1.0265     1.061
trt1    0.8641     1.1572    0.6167     1.211

Concordance= 0.629  (se = 0.024 )
Likelihood ratio test= 33.25  on 3 df,   p=3e-07
Wald test            = 34.87  on 3 df,   p=1e-07
Score (logrank) test = 35.31  on 3 df,   p=1e-07
```

If the model does not converge i.e shows <Inf> for a variable or group in the confidence interval, or the hazard plot cannot be displayed, it means that there is not enough data in that category to calculate hazard ratio. Consider filtering your data or using different covariates.

Figure 17. Summary of the Cox PH model.



For categorical variables, the reference group is set at a category, or at the lowest value if the groups have numerical coding. The reference group is used for comparison and is set at a hazard ratio of 1. With a continuous variable, the hazard ratio indicates the change in the risk of event if the parameter in question rises by one unit. A hazard ratio of 1 means lack of association, a hazard ratio greater than 1 suggests an increased risk, and a hazard ratio below 1 suggests a smaller risk.

If the model does not converge i.e shows <Inf> for a variable or group in the confidence interval, or the hazard plot cannot be displayed, it means that there is not enough data in that category to calculate hazard ratio. Consider filtering your data or using different covariates.

Figure 18. Forest plot of the Cox PH model.

12. Documentation

The last tab contains documentation and instructions how to use ClinFlow.

4.2. Limitations and future work of ClinFlow

The application's current version can be viewed as a proof-of-concept. ClinFlow requires extensive validation and usability testing before actual deployment. The list below shows some limitations identified in the application.

1) Scalability. The application requires extensive optimization and analysis of response times when used with large datasets. Performance requirements for this application have not been set at the moment.
2) Generalizability. The data wrangling and analysis options in the application it do not cover the complex challenges that are specific to the domain and data collection practices of each clinical dataset. The application has been designed for longitudinal follow up data and survival data.
    a) If the dataset contains mostly categorical variables, PCA, T-SNE, MDS, SOM and K-Means clustering might not be the best approaches to analyze this type of data.

b) The missing data and outlier treatment options are limited, and they are based on visual exploration. More complex outlier detection or data imputation methods have not been included.

c) The survival analysis can be performed only with baseline covariates and does not include time-adjusted covariates for the Cox proportional hazards model.

Thorough validation and usability testing have been left for the future. The application can be locally hosted, making it safe to use from the data privacy point of view. Future work also includes experiments with different datasets and increasing the generalizability to fit more types of clinical data. Some improvement ideas for the functionalities of the application that could be implemented in the future are listed below.

1) Including a general preprocessing pipeline along with user options for the more complex preprocessing operations.
2) Including various advanced methods for outlier detection and missing data imputation
3) Including options to create new features in the data using complex formulas that include multiple variables, not just one variable at a time.
4) Implementing supervised learning methods such as regression or classification models along with the visualizations present in the app.

4.3. Example study

For demonstration, we used our tool to inspect the DIPP data [3] from Oulu University Hospital and to study the relationships between islet autoantibodies and progression to T1D. We compared our results with the studies of Knip et al [4] and Pöllanen et al [5] that demonstrate how different types of autoantibodies, autoantibody combinations and age at seroconversion are associated with disease progression.

Before the analysis, a few more complex preprocessing steps were needed in order to bring all the data into a structured longitudinal follow up format, and extract important information in the form of new features. These steps are too complex and dataset-specific, so they were performed before the ClinFlow application was used.

Preprocessing

To obtain a longitudinal follow-up dataset, data transformation and feature construction operations have been applied on the cleaned data, based on literature study and domain knowledge.

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

Data transformation

In this stage, we converted character variables that are difficult to read into categorical variables that are easier to read. Depending on the variable and considering the domain literature, appropriate categorization of the variable has been done, while keeping the original features as well. The following list describes each transformation applied to the DIPP variables.

1) Autoantibody values have threshholds for positive/negative values but they are not obvious in the dataset.
   a) Created factor variables for each autoantibody TRUE/FALSE if the value goes over the positive threshold.
2) Pregnancy duration column is in character form, with duration as weeks + days.
   a) We converted the variable "duration" into a factor column with three levels: “premature”, ”normal”, “prolonged”.
   b) We performed text parsing for the number of weeks (<37 - premature, >41 - prolonged).
3) The month of birth might have influence on diabetes risk [37].
   a) We created month of birth factor column “birth_month” from the birth date.
4) Pets variable is in free text format.
   a) We created a visit variable “pets” factor with three levels: “cat/dog”, “other”, “no pets”.
   b) If in the type_of_pets we find the following string patterns : “koir” or “kiss”, we assign factor “cat/dog” (because in this dataset, most of them appear together in the same entry).
   c) If in the type_of_pets there is missing value or empty string “” and in the variable “is_pets” is 0, we assign “no pets”.
   d) If in the type_of_pets there is anything other than the text patterns “koir” or “kiss”, including missing value or empty string “” and in the variable “is_pets” is 1, we assign “other”.
   e) If in the variable “is_pets” is value 2 or missing and in the type_of_pets is missing or empty string “”, we assign NA (we don’t know if there is a pet or not) [38].

© The HTx Consortium 2019-2023. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement Nº 825162.

5) Dataset contains children monitored from birth, their parents and their siblings in the same set, with different letters in the patient code but no other clear distinction between them.

   a) We created a variable named “who” with four categories: “child” - for children monitored since birth, “mother”, “father”, and “sibling” - for the siblings of children monitored from birth, by text parsing the patient code. Here, we know that the children monitored from birth are most likely to have the information from early life.

Feature construction

This stage is the most important and complex. While the tool also enables some user-defined feature construction operations, we have constructed some ready-made features that are relevant for the DIPP study based on domain literature. Some of them are derived from variables that are most prevalent in clinical data, but most of them are specific to the DIPP study. Some of these feature construction techniques can be later generalized to other datasets of the same type. For the new more complex variables that consider autoantibody values, we need to take into account and exclude from calculations the positive autoantibodies transferred from the mother or present in the cord blood. For this, we have assumed that a positive sample has autoantibodies from the mother or cord blood if it meets the following criteria:

1. The child’s mother is present in the dataset and has positive autoantibodies OR the child’s cord blood sample is present in the dataset and is positive.

2. The sample is from before the age of one year old.

3. The sample is not preceded by a negative sample.

We consider positive, any sample that has one or more positive autoantibodies. We created a variable called "mother_samp" that marks TRUE for the samples meeting the above criteria, and FALSE for all the other samples. A logical scheme of this process is presented in the flowchart in Figure 19.

© The HTx Consortium 2019-2023. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement Nº 825162.

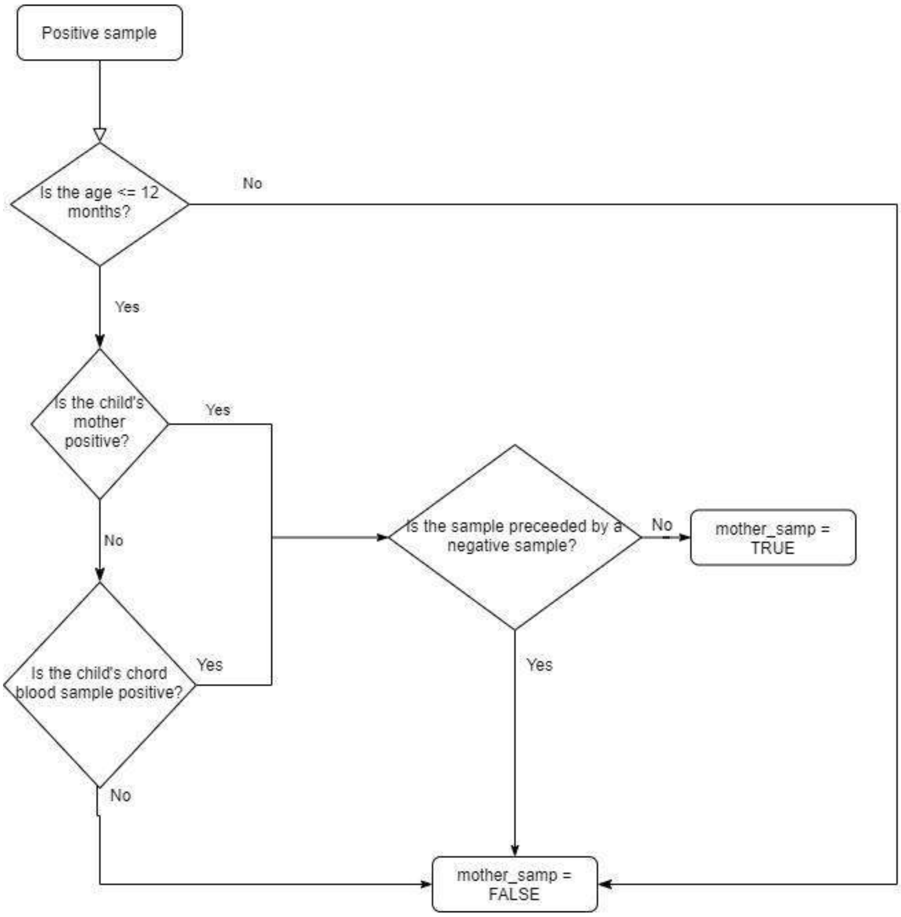

Positive sample
Is the age <= 12 months?
No
Yes
Is the child's mother positive?
Yes
No
Is the sample preceeded by a negative sample?
No
mother_samp = TRUE
Yes
Is the child's chord blood sample positive?
Yes
No
mother_samp = FALSE

Figure 19. Flowchart of detection of samples with positive autoantibodies transferred from the mother.

The following list describes each feature construction technique applied on the DIPP data.

1) Age at the visit is not present in the data.

   a) We calculated age at that visit with birth date and visit date.

2) Maximum follow up time is important. Data should be compared from patients with similar follow up times.

   a) We created a new variable called "Age_follow_up" that marks the maximum visit age of each patient.

3) Disease progression time is important [5].

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

   a) We created a variable "progression_time that marks the time passed between the seroconversion and the diagnosis dates. For the patients who never got diagnosed with T1D the progression_time is NA.

4) Age of the mother at birth could have an impact on analysis [39].

   a) We calculated age of mother at birth from the mother's birth date and her (corresponding patient code) child birth date, and created a variable "Mom_birth_age"

5) Blood sample from a visit is positive if any >=1 of the autoantibodies is positive.

   a) We created a binary factor visit variable "pos_sample" to mark if the sample(row) is positive in any >=1 antibody.

6) Two consecutive positive samples in any >= 1 antibody for a patient is considered true positivity [5]. For each individual patient code:

   a) We excluded positive samples that were identified as transferred from the mother.

   b) If at any point there are two or more consecutive TRUE visit rows in any of the four columns with antibodies, we assign a factor variable "POS_antibodies" = TRUE for that patient.

7) Age of seroconversion has an importance in analyses [5]. We found the first positive sample in any >=1 antibodies and retrieved age from that visit as follows:

   a) We excluded positive samples that were identified as transferred from the mother.

   b) We created a variable called "Age_seroconv" that contains the age of the first positive sample.

   c) For the patients who never had a positive antibody sample, Age_seroconv is NA.

8) Seroconversion type is important [5]. Are multiple autoantibodies present at seroconversion and which ones?

   a) We created a factor variable to mark whether only one antibody was positive at seroconversion, or multiple, and whether it was IAA or other[79]. Factor levels are: "ICA","GADA","IAA","IA2A","multipositivity IAA" for the ones with IAA combined with other autoantibodies at seroconversion, and "multipositivity" for the ones with 2+ positive antibodies at seroconversion but no IAA.

9) The type of positivity after seroconversion is important [5].

    a) We excluded positive samples that were identified as transferred from the mother.

    b) We created a factor variable "positivity_type" with three levels: "single" for patients with at least two consecutive positive samples in only one autoantibody at a time, "multi" for patients with at least two consecutive positive samples in two or more autoantibodies in the same time, and "negative" for patients with one or zero positive samples in any autoantibodies.

10) Antibody titre at seroconversion for each autoantibody is important [5].

    a) We retrieved the antibody value from visit rows where visit age = seroconversion age.

    b) We created a new titre variable for each autoantibody that contains the value of that autoantibody at seroconversion, regardless of the seroconversion type.

    c) Patients who never seroconverted have NA in these variables.

The final dataset used for analysis has a total of 54 attributes.

Analysis with ClinFlow

The preprocessed longitudinal follow up data is analyzed with ClinFlow, for the purpose of demonstrating the capabilities of this tool, and showing how we can obtain results that are in line with other articles part of the DIPP study.

When data is uploaded into the tool, the constant and time-series variables are automatically detected, and data is split into two tables: *patient data* and *visit data*. Visit variables can be summarized over a time period and added into the *patient data* table - this allows the user to check the impact that values from medical visits from a certain time interval have on an outcome, or compare the values of different groups, even if the subjects had irregular medical visits.

We used the tool to identify a subset of 253 children with persistent positivity (two or more consecutive positive samples) in at least one autoantibody. 118 of them progressed to T1D before the age of 15 years. The rest were followed up until 15 years old and did not progress to T1D. The subset was created using the data filter, to include children monitored from birth until at least 15 years old, who seroconverted after the age of one year to at least one autoantibody. We excluded the children under 15 years old without a T1D diagnosis and the children who seroconverted before the age of 1 year. The filter option of the tool made it easy to group the

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

data and subset it using the dynamic user inputs (sliders and drop-down lists), and check the summary statistics of the groups. We performed clustering and exploratory analysis to identify the factors associated with the progression to T1D for this subset. We also created a structured panel dataset using the *visit data* by aggregating visit values from certain time intervals to explore the time trends of these values.

Additionally, we performed Cox regression survival analysis to visualize the time-to-diagnosis hazard ratios for T1D diagnosis.

4.4. Results

Using the tool, we were able to identify several associations in the data, Figure 20 shows illustrations directly exported from ClinFlow. We performed PCA and visualized the clusters formed by plotting the principal components (PC). We found a strong relationship between multipositivity (positive samples of two or more autoantibodies occurring at the same time) and progression to T1D, and no relation between single positivity (positive samples of only one autoantibody) and progression to T1D. Figure 20A illustrates the PC values. The points are shaped according to diabetes progression (squares stand for progressors and circles for non-progressors) and the color indicates positivity type (red for multipositivity, yellow for single positivity). An association was found between multiple seroconversion at an early age, and progression to persistent multipositivity and T1D. Figure 20B shows the relationship between multipositivity (circles) and multiple seroconversion (green).

Figure 20C shows a heatmap of loading scores that explain how much each variable contributes to the variation in the data, with seroconversion type contributing the most, followed by the age at seroconversion.

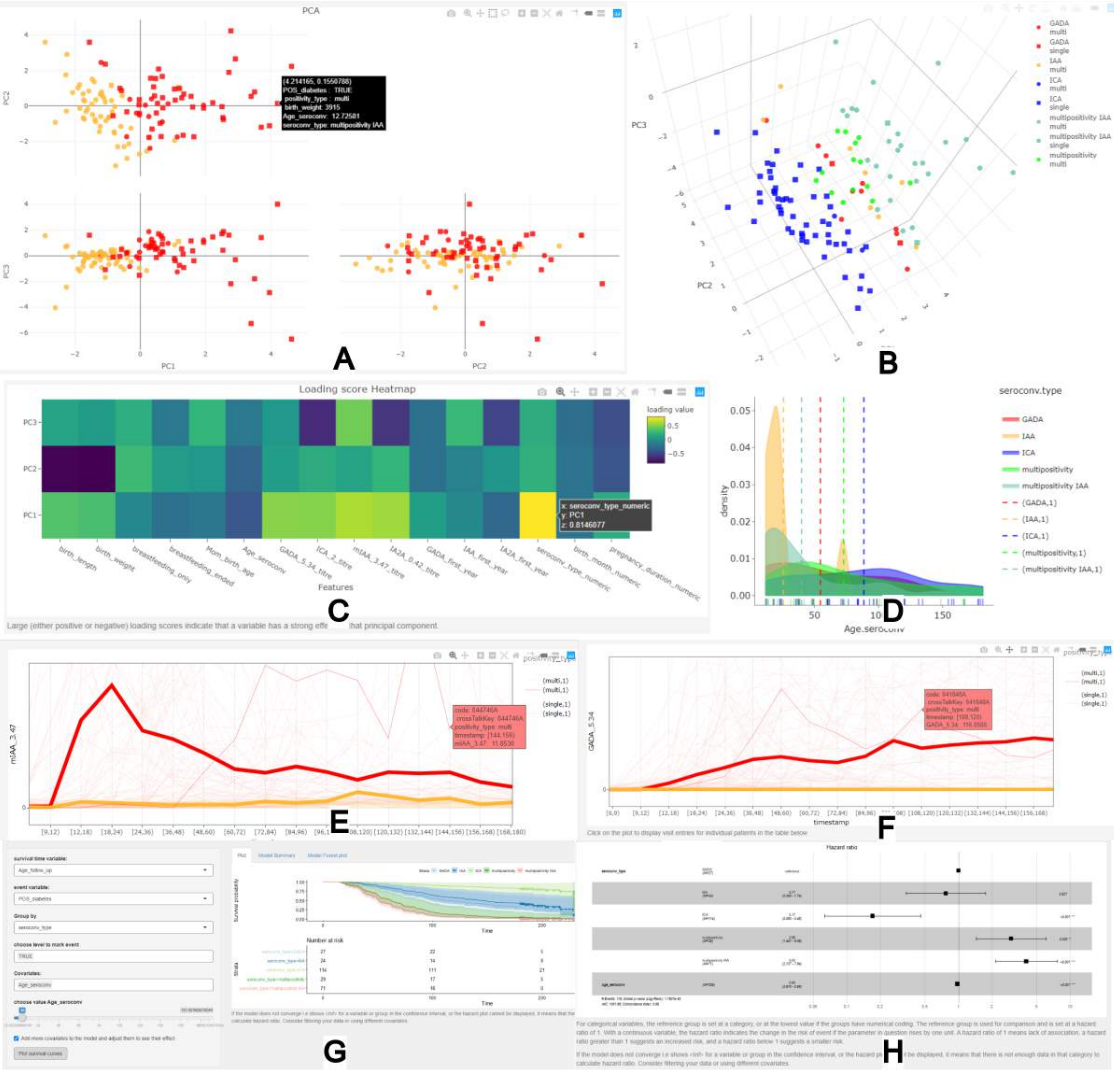
The IAA autoantibody positivity is associated with early seroconversion. In Figure 20D, the density distribution of the seroconversion age shows a peak around the age of 2 years for the groups with IAA at seroconversion.

Higher values of IAA autoantibody in early life and high values of GADA later in life are indicative of progression to multipositivity, according to the time trend plots in Figures 20E and 20F.

The survival analysis results confirm these findings in Figure 19G. Additionally, they show an association between multiple seroconversion and an increased risk of rapid progression to T1D.

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

The Kaplan-Meier curves in Figure 19G show that multiple seroconversion has the lowest time-to-diagnosis.



Large (either positive or negative) loading scores indicate that a variable has a strong effe[...]hat principal component.

Click on the plot to display visit entries for individual patients in the table below

Figure 20. 2D PCA clustering (A). 3D PCA clustering (B). Heatmap of loading scores (C). Density plot of seroconversion age for each autoantibody (D). Time trend of IAA values (E) and GADA values (F) for multi-positive (red) and single positivity (yellow). Survival curves, where time is the subject’s age in months (G).

© The HTx Consortium 2019-2023. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement Nº 825162.

These findings are in line with the results of Knip et al [4] and Pöllanen et al [5], and show how the power of ClinFlow, and it can help and speed up analysis of medical data. Therefore, ClinFlow is a helpful tool to illustrate and understand the data structure. The next chapter shows how this understanding can be used to create synthetic datasets.

5. Understanding the structure of the data to create synthesized data

5.1 Introduction

The maintenance of data obtained as a part of the clinical analysis has evolved. Initially, medical data was generated and maintained by health care professionals in the form of several Electronic Health Records (EHR). Nowadays, most countries possess a centralized EHR system to accommodate the availability and completeness of the data. The purpose of centralized EHR is to hold detailed information about the patient's medical archives in one place. These centralized EHR can later be combined with other data sets to help medical professionals administer the best possible treatment with knowledge gained from data by using next-generation technologies, including artificially intelligent systems, and potentially transform medicine. Despite the benefits, few considerable obstacles prevail in the process of exploring and achieving this goal [6]. Some are associated with the content and structure of the modern healthcare database; others regard the complications and expense of producing and sustaining comprehensive databases. However, those rich veins of data are too often locked away. There are several reasons why every data set is not publicly available or why it is not even possible to share the data within university or project partners. Most often, data collectors do not get the recognition of their investment in data collection; hence the desire to be the first to explore and utilize the data before they sell or distribute it to others. Nevertheless, one being the most important reason is the privacy of the subjects.

Clinical data either collected as a part of research or recorded during clinical practice is reckoned confidential and required to be pseudonymized or anonymized before leaving the hospital [7]. Pseudonymization and anonymization techniques consist of altering and removing explicit identifiers such as names, addresses, and national identity numbers from a dataset. However, in pseudonymization, a person can still be re-identified by data linking, leading to a reduction in k-anonymity [8, 9], and anonymization techniques have failed multiple times in the past [10]. To further reduce the risk of re-identification, data scientists use data aggregation techniques and

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

induce random noise to the data, which leads to distortion of the relationship between variables in the data set [8]. Having a data set with distorted relationships between variables can be misleading. As the correlation between the feature's changes, the risk that correlation is interpreted as causation increases and can lead to misconceptions.

The generation of the synthetic data set, which preserves the statistical properties of the original data set and, simultaneously, ensures the patient’s privacy, will be the fittest case in the current scenario to share data. A data synthesis tool termed Synthpop was explored and examined while underlining the statistical properties, machine learning applicability, and quality of information contained in the data set to produce synthetic dataset. The primary objective was to question the performance of the synthesis tool by evaluating the impact of data synthesis procedure; Over two different data sets for comprehensiveness evaluation. The data set is the Finnish Type 1 Diabetes Prediction and Prevention (DIPP) study database [3]. Impacts of data synthesis was measured based on the general and specific utility, and quality of information of the synthetic data set compared to the original data set. General utility measures will evaluate the difference in the statistical properties of the data sets, and specific utility measure will focus on the performance of the fitted models over different data sets (synthetic and original). One null and one alternative hypothesis was defined, evaluating the difference in the results of utility measures. The Synthpop will succeed in a performed test, if results fail to reject the null hypothesis, which states that the two data sets (synthetic and original) have at most a statistically non-significant difference. Moreover, the study will be finalized via evaluating from an information-theoretic point of view, by analyzing entropy and mutual information within the data sets or in comparison to measure the quality of information contained in the data set.

5.2 Methodology

Synthpop

Data synthesis is a process of generating data that mimics the original data set but does not hold any disclosure records. Figure 21 represents the workflow of generating synthetic data in brief and Figure 22 gives details of sub-processes.

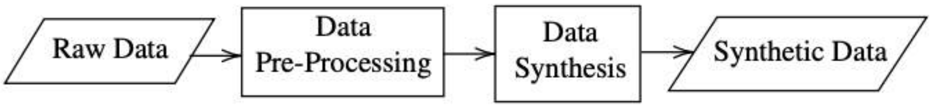


Figure 21. Workflow of generating synthetic data.

© The HTx Consortium 2019-2023. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement Nº 825162.

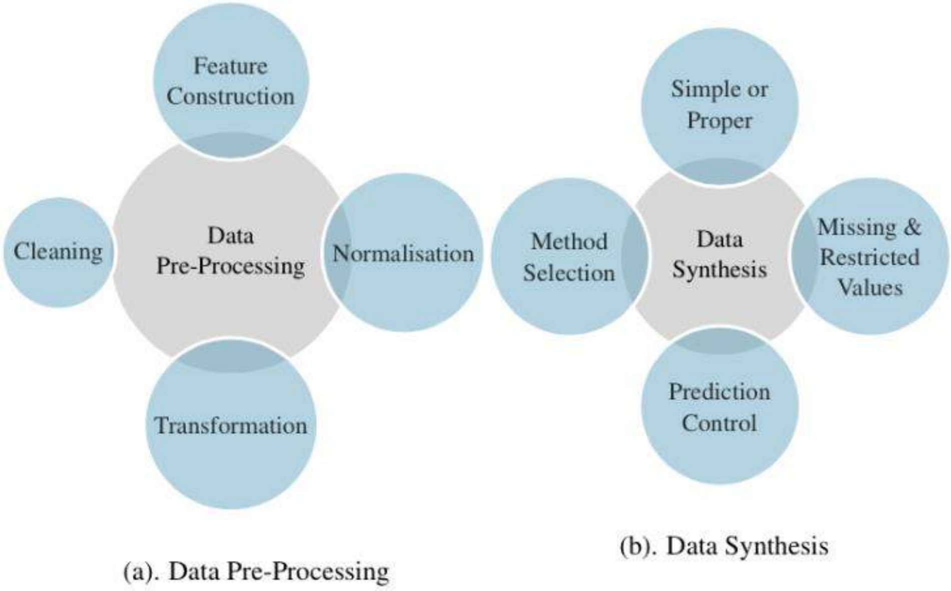


Figure 22. Data pre-processing and synthesis.

The tool for data synthesis used in this research is an R package termed synthpop [12]. The synthpop package was written as a part of the Synthetic Data Estimation for UK Longitudinal Studies (SYLLS) project. Formerly to share the sensitive population-level data outside the setting where researchers were holding the original data set. Later, the synthpop package was altered to makes it applicable to other data sets.

The method works by replacing some or all observed values by sampling from an appropriate probability distribution, conditional on the variable to be synthesized, the values from all previously synthesized columns of the original data set, and the fitted parameters of the conditional distribution (simple synthesis) or posterior predictive distribution of parameters (proper synthesis) while retaining the statistical properties of the original data set and relationships between the variables. The synthetic data can be produced simply via *syn()* in a single command providing a data set, which is a data frame or matrix to be synthesized. Users can customize the synthesis of a data set according to requirement, applicability, and type of data variables for better performance of the overall system. By default, the *syn()* function produces one synthetic data set, but multiple data sets can be generated by setting the parameter *m* to a coveted number. An additional parameter *seed* can be used to fix te pseudo-random number generator to reproduce the same results. By default, *syn()* function uses simple synthesis but proper synthesis can be done by setting the proper argument to *TRUE.*

Methods for Synthesis

The Synthpop consists of both parametric and non-parametric methods. Each method generates synthetic values for each variable sequentially. Synthetic values are generated using the distribution of variable to be synthesized conditional on the distribution of previously observed synthetic and original variables called predictors. The default method of synthesis is *"cart"* for all variables with predictors. The method *"cart"* is a non- parametric method based on Classification and Regression Tree (CART); capable of handling any type of data. However, the first variable to be synthesized in the data set does not have a predictor, and it is a particular case where its values are by default generated by random sampling with replacement from original values (*"sample"* method). However, the user does not need to use the same method of synthesis for all variables with predictors; a user can assign different methods from the list of methods to each variable in the data set befitting the type of data. On the other hand, by setting parameter method to *"parametric"* assigns default parametric methods to each variable based on their data type. Furthermore, if a user does not want to change or synthesize a variable, an empty method (" ") should be used for that variable. Finally, a new method of synthesis can be defined by writing a function named *syn.newmethod()* and for synthesis, specify the method parameter of as *"newmethod".*

Controlling the Sequence and Prediction

Synthetic values of each variable are generated from a joint distribution. The joint distribution is defined in terms of a series of conditional distributions. The values are imputed sequentially from the distribution of the variable to be synthesized conditional on two distributions: 1) The distribution of all previously observed variables in the original data set, 2) The distribution of all previously synthesized variables. This sequential process is by default automated, following the order of how variables appear in the data set (left to right). However, the order can be changed or specified for each variable by listing out the indices of columns in the desired order to set parameter *visit.sequence*. If a user wishes not to synthesize a variable and not use it as a predictor, it should be removed from the *visit.sequence*. Furthermore, if a user wishes not to synthesize a variable, yet wishes to use the variable as one of the predictors for the synthesizing model, then an empty (" ") method should be used while keeping the variable in *visit.sequence*. Note that variable/s to be synthesized later in *visit.sequence* cannot be used as predictor/s for variable/s which appears before it. Though, variable/s can explicitly be removed as a predictor/s for any specific variable/s by updating the *predictor.matrix*. The *predictor.matrix* is a matrix with ones and zeros; Ones indicates that the variables should be used in the prediction model for generating synthetic values for a particular variable and zeros for otherwise.

Handling Data with Restricted or Missing Values

Relationship between variables can diversify significantly within a data set. Some variables can have a dependency on each other or could be tightly linked. As the goal of the synthetic data is to mimic all characteristics of the original data, these restrictions should be preserved during the data synthesis process. For example, in a medical data set, the variable containing information about the patient's sibling's medical history is restricted to the variable containing information whether the patient has siblings; This restriction needs to be addressed in order to get the best results out of the synthesis process. Simply when other variables determine the value for some case, the rule and corresponding values should be specified using rule and *rvalues* parameters. Furthermore, if the data set has missing values and the values are defined with something distinct than the R missing data code *NA*, it should be specified in cont.na parameter of the *syn()* function. Missing values in categorical variables are handled as additional categories. However, missing values in continuous variables are modelled in two steps. First, an auxiliary binary variable is synthesized to model whether a value is missing or not, and if there are multiple types of missing values, an auxiliary categorical variable is created to record this. Second, a synthetic model is fitted to non-missing values, and synthetic values are generated for non- missing categories in the auxiliary variable. Finally, the auxiliary variable, variable with non-missing values, and zeros for remaining records are used for prediction of other variables.

Utility Measures of Data

The purpose of a synthetic data set is to resemble all the properties of the original data set. Thus, analyses made on synthetic data set should lead to the same conclusions to the analyses made on the original data set. In theory, to achieve the formally mention purpose, the model used for the synthesis process should resemble the process of the original data generation. The methods to assess the utility of the synthetic data set can be broadly divided into two approaches: general utility and specific utility [13]. General utility assesses whether synthetic data have overall similarities in the statistical properties and multivariate relationships with the original data set. Whereas specific utility assesses the similarity of performance of a fitted model on the synthetic data to the original data. The **Synthpop** package provides two types of analyses for the synthetic data set based on the general and specific utility of the data set utilizing the *compare()* function in the package. First is the relative frequency distribution, and second is the

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

linear machine learning model’s confidence interval overlap. However, in this study, besides relative frequency distribution from the package, more rigorous analyses will be performed.

The overall utility of the synthetic data will be assessed on how adequately synthetic data succeed at all conducted utility tests. In order to succeed at a utility test, synthetic data need to resemble all the properties of the original data with at most statistically non-significant difference. For formal assessments, hypothesizes will be as follows: Let *D* denote an original data set, and S_i denote a synthetic data set where *i* indicates the index for synthetic data produced with the different synthesizing method. Let *t* denote a vector of tests which returns a statistic, and *C** be a comparison function which returns a *p-value*. Finally, comparing the output of *C** with *α*, a threshold value for the level of significance.

$$H_o : C^*\{t(D), t(S_i)\} \geqslant \alpha, \qquad \text{for all } t \in [0, \tau]$$

$$H_a : C^*\{t(D), t(S_i)\} < \alpha, \qquad \text{for any } t \in [0, \tau]$$

The quality of the synthetic data will be estimated based on whether utility tests lead to failing to reject the null hypothesis. In order to fail to reject the null hypothesis, synthetic data must have *p-value* larger or equal to *α* for all utility tests. The null hypothesis will be rejected if synthetic data possess p value smaller than *α* for any utility test leading to accept the alternate hypothesis. Note that the *α* is set to 0,05 for all tests.

General Utility Measures

A visual representation of data helps data analysts to process information from data faster than from written information. Visualization of frequency distribution can reveal a lot about the data and its properties. Four principal characteristics of the frequency distribution are:

1. The measure of central tendency and location (mean, median, mode)

2. The measure of dispersion (range, variance, standard deviation)

3. The extent of symmetry/asymmetry (skewness)

4. The flatness or peakedness (kurtosis)

On the other hand, relative frequency distribution provides the fraction or proportion of times a value occurs in data sets. A side-by-side univariate distribution of each variable in the synthetic and original data set was plotted to compare the changes in the probability distribution, which can be used to determine the likelihood of specific results to occur within a given population. Furthermore, the two-sample Kolmogorov–Smirnov test was used to evaluate whether two

© The HTx Consortium 2019-2023. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement Nº 825162.

underlying one-dimensional probability distribution differs in two different data sets (original and synthetic data set) for each variable. Apart from visualizing frequency distributions, visualization of data points itself can help data analyst have a look at data from a different perspective. Visualization of data directly, which has more than three dimensions is currently out of scope, but dimension reduction techniques which preserve the relationship between variables can be used. Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for data visualization similarly to T-distributed Stochastic Neighbor Embedding (t-SNE) [14], but also for general non- linear dimension reduction [15]. Finally, the bivariate Pearson Product-Moment Correlation Coefficient (PPMCC) was used to measure the linear correlation between pairs of continuous variables.

Specific Utility Measures

The specific utility of the data can be assessed by comparing the performance of the fitted synthetic and original models. In this study, multiple machine learning models were used as classifiers, such as Gradient Boosting Machine [16, 17, 18], Pattern Recognition Network [19], k-Nearest Neighbours [20], and Linear Discriminant Analysis [21]. Most of these models can also be used for regression. Different types of machine learning models were used to evaluate the generality of the primary method of synthesis "Synthpop". Moreover, the performance of the fitted model will be examined on multiple parameters for overall performance estimation.

Quality of Information Content

The goal of data anonymization procedure is to reduce semantics, meaning minimizing or removing personal information in a data set [22, 23]. Data anonymization can cause distortion and information loss in the data set. In this section, the concepts of information theory were used, to quantify the level of distortion and information loss. Concepts such as evaluating change in entropy and estimating the mutual information (MI) between variables was used [23, 24].

Entropy

Entropy is a fundamental quantity in information theory associated with any random variable. Entropy can be interpreted as the level of information, surprise, or uncertainty associated with the value of a random variable or the result of a random process [25]. The bit, which is the unit of entropy, is adopted as a quantitative measure of information, or measure of surprise. The entropy of a random variable *X*, with possible outcomes X_i, each with a probability of occurrence $P_x(x_i)$ is calculated as:

$$H(X) = -\sum_{i} P_X(x_i) log_b P_X(x_i)$$

The entropy is maximum when all outcomes are equally likely in a system. If the system moves away from equally likely outcomes or introduces some predictability, the entropy goes down. The fundamental idea of the information theory is that, if the entropy of an information source or system or data set drops, that means fewer questions are needed to ask to guess the outcome. Entropy is directly proportional to uncertainty, i.e., as the value of entropy increases due to unpredictability, uncertainty in the system's outcome increases and the ability to compress decreases, similarly, if the value for entropy decreases due to known structure, then the ability to compress increases, which lead to entropy being indirectly proportional to the ability to compress.

Mutual Information

The MI is a measure of mutual dependence between two random variables. MI measures the information gain for a random variable *X* when information about another random variable *Y* is given. MI between two random variables *X* and *Y* can be calculated as:

$$I(X,Y) = \sum_{x_i \in X, y_i \in Y} p(x_i, y_i) log(\frac{p(x_i, y_i)}{p(x_i)p(y_i)})$$

or

$$I(X;Y) = H(Y) - H(Y|X)$$

If entropy *H(Y)* is a measure of uncertainty about a random variable Y, then *H(Y|X)* is a measure of what *X* does not say about *Y*. In other words, *H(Y|X)* is the amount of uncertainty remaining about *Y* after *X* is known. Therefore, the equation can be interpreted as the amount of uncertainty in *Y*, minus the amount of uncertainty in *Y* after *X* is known. Furthermore, this provides the inherent meaning of MI as the amount of information or reduction in uncertainty that one random variable provides about the other.

The Kraskov's estimator [24] of mutual information is closely related to Shannon's entropy, but Kraskov's estimator relies on the count of nearest neighbors. Kraskov's estimator, along with many others, uses canonical distance defined in metric space for computability over Euclidean space and uses Euclidian distance as the distance function. The mutual information estimator $I^{(2)}$ between two random variables x_i and y_i is defined as:

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

$$I^{(2)}(X,Y) = \Psi(k) - 1/k - < \Psi(\mathbf{n}_x) + \Psi(\mathbf{n}_y) > + \Psi(N),$$

with the digamma function and k denoted the number of neighbors. Where < · · · > denotes the averages of both vectors $n_x(i)$ and $n_y(i)$ holding counts of neighbors over all $I^{(2)}$ *[1,. . . ,N]* and over all realizations of the random samples.

In this study, a variation of the second algorithm from Kraskov's estimator proposed by Oliver et al. [23] to use the method over non-Euclidean spaces using non-Euclidean distances will be used. Where the calculation requires the nearest neighbors of points in joint space and counting how many lies in an absolute ball.

5.3 Experiments and results

The data set used in this study was built and pre-processed from the original DIPP database and modelled using ClinFlow.

Preprocessing

The data until the age of 12 months was aggregated to utilize information gain from that data to predict the positivity of the autoantibodies later in life. First, variables such as infections were aggregated to value 0 if the number of infections is zero or to value 1 if more than one or two infections in the first 12 months of age. Infections leading to hospital care and other similar variable were cumulated similarly. Furthermore, for variables such as autoantibodies, the maximum autoantibody value was considered before the first positive value before 12 months of age occurred. Later excluded the seven subjects whose autoantibodies values were in positive range before 12 months of age due to autoantibodies transmitted from mother. Finally, a response variable "POS_antibodies" was defined based on the positivity of autoantibodies. Class negative, if the subject never had an occurrence of positive value in any autoantibodies up until 170 months of age and class positive, if the subject had two or more consecutive positive value occurrences in any autoantibodies up until 170 months of age. The value of an autoantibody is positive if they are higher than a specific threshold for the respective autoantibodies. The threshold values for GADA, IA2A, and IAA are 5.34, 0.42, and 3.47, respectively. Overall, providing 30 attributes using a small subset of data of 1329 subjects. Out of which 839 subjects belong to the positive class and 28 490 to the negative class. Table 1 provides the list of all attributes in the data set and their description. The goal of the data set is to predict the probability of the positivity of autoantibodies before the age of 15 years by utilizing information gain from the first 12 months of data.

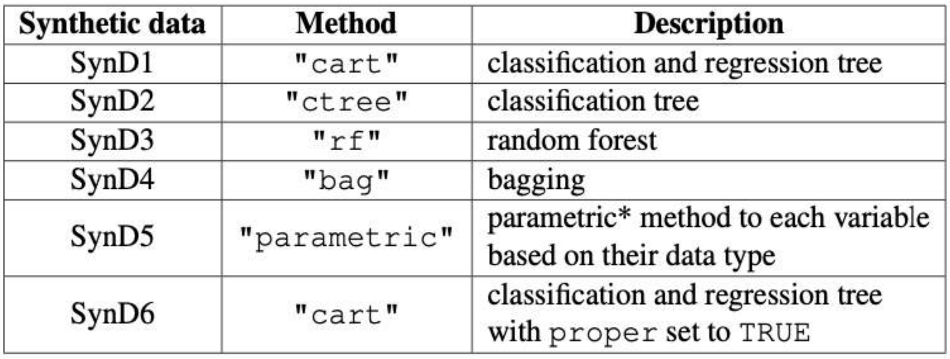
Experiments

The pre-processed version of the DIPP data set is a data frame with 30 attributes, including the response variable for 1329 subjects in total. Multiple variables in the data set were first turned into factors using *as.factor*() command in R. Later, the data set was synthesized numerous times via *syn()* command from Synthpop package using several methods. As mentioned earlier in Subsection 3.1.2, the first variable to be synthesized in the data is by default generated using *"sample"* method. In our case, the response variable *"POS_antibodies"* is the first variable to be synthesized, and then the rest of the attributes. Table 3 provides the list of all attributes in the data set, and their description in the order of synthesis, i.e., *"visit.sequence"*.

Methods of synthesis

Both non-parametric and parametric methods of synthesis were used in the engendering of the synthetic data sets. Table 2 lists the methods used for generating the corresponding synthetic data with denoting names and method description. For synthesis of *SynD5*, *"parametric"* method was applied. For all non-parametric method, every attribute was synthesized using same method. However, for SynD5, different parametric methods were applied depending upon the type of the attribute. Each synthetic data set was generated using seed value for result replication. A total of 5 synthetic data sets were generated for initial experimentation using 5 different methods *(SynD1 to SynD5)*. One method of synthesis which performs the best out of 34 those 5 methods was selected for generating another synthetic data set by setting the argument proper to TRUE for proper synthesis for further analysis *(SynD6)*.

Table 2. Denoted names for synthetic data sets and methods used for creation, *List of parametric method for each variable is listed in Table 2.



Synthetic data	Method	Description
SynD1	`"cart"`	classification and regression tree
SynD2	`"ctree"`	classification tree
SynD3	`"rf"`	random forest
SynD4	`"bag"`	bagging
SynD5	`"parametric"`	parametric* method to each variable based on their data type
SynD6	`"cart"`	classification and regression tree with `proper` set to `TRUE`

Specific Utility

In this section, we evaluated whether different methods of synthesis preserve the specific utility of the original data set differently, after which we selected one method of synthesis that performs best out of all methods used. The goal is twofold, first to investigate if synthetic data sets can be used for machine learning problems when the original data cannot be acquired and second to assess how well synthetic data sets perform on the machine learning classifier as compared to the original data set.

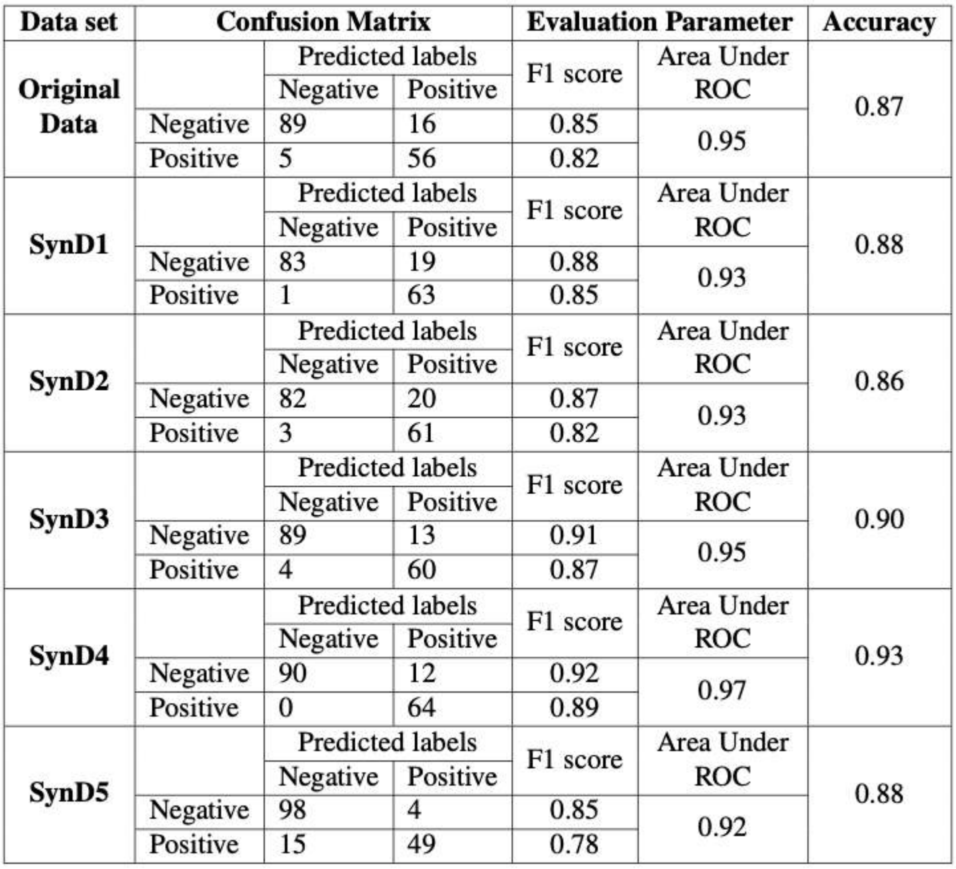
The machine learning classifier used is the GBM model, which was fitted, validated, and tested 10 times (for more stable performance of the model) with all data sets from Table 1 along with the original data set, each time with different seed value. Additionally, each data set was divided into three splits with different seed value, before model fitting, 75.0% of data for training, 12.5% for validation, and 12.5% for testing.

Comparing different methods

We compared the results obtained from synthetic data test sets to the results of the original data test set; to evaluate which synthesizing method produces the synthetic data set principally resembling the performance of the original data set. The performance measure used is confusion matrix and parameters derived from it. The motivation behind using multiple performances evaluate parameters is to provide a more robust interpretation, as a model can have very high accuracy, yet suffer from low precision.

One sample out of ten for the comparative performance of synthetic data sets for all selected synthesis methods with the original data set can be seen in Table 3. Note that the process was repeated 10 times for all data sets to perform a significance test over testing accuracies. The accuracies of each synthetic data set fitted model and the accuracies of the original data set fitted model were compared using *C** comparison function. The *C** function returns a *p-value*, Table 4 provides the *p-value* for each data set. Every single p value was calculated using t-test, comparing accuracies of every synthetic data set to the original data set over the GBM model, 10 iterations.

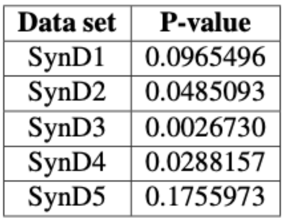
Table 3. Data sets and their performance over GBM model.



Data set	Confusion Matrix			Evaluation Parameter		Accuracy
Original Data		Predicted labels		F1 score	Area Under ROC	0.87
		Negative	Positive			
	Negative	89	16	0.85	0.95	
	Positive	5	56	0.82		
SynD1		Predicted labels		F1 score	Area Under ROC	0.88
		Negative	Positive			
	Negative	83	19	0.88	0.93	
	Positive	1	63	0.85		
SynD2		Predicted labels		F1 score	Area Under ROC	0.86
		Negative	Positive			
	Negative	82	20	0.87	0.93	
	Positive	3	61	0.82		
SynD3		Predicted labels		F1 score	Area Under ROC	0.90
		Negative	Positive			
	Negative	89	13	0.91	0.95	
	Positive	4	60	0.87		
SynD4		Predicted labels		F1 score	Area Under ROC	0.93
		Negative	Positive			
	Negative	90	12	0.92	0.97	
	Positive	0	64	0.89		
SynD5		Predicted labels		F1 score	Area Under ROC	0.88
		Negative	Positive			
	Negative	98	4	0.85	0.92	
	Positive	15	49	0.78		

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

Table 4. Data set and p-values for their comparative accuracy with original data over GBM model (fitted 10 times)



Data set	P-value
SynD1	0.0965496
SynD2	0.0485093
SynD3	0.0026730
SynD4	0.0288157
SynD5	0.1755973

.

The objective is to fail to reject the null hypothesis, i.e., the difference in the performance of the synthetic data should differ with the performance of the original data with at most non-significant difference. In other words, aiming that the synthetic data set produced using any method does not need to perform better or should not perform worse than the original data, but it needs to perform as close as possible to the original data set. From Table 10, the data sets produced using method *"cart" (SynD1)* and "parametric" *(SynD5)* are only two data sets with p value greater than α, whereas rest have p value smaller than α. If the p value is greater than α, it states that according to statistics, these two data sets fail to reject the null hypothesis meaning that difference is statistically non-significant. Moreover, from Table 3, we can say that *SynD1* performs better than *SynD5* when other evaluation parameters are considered. As the overall performance difference to original data is smaller for *SynD1* as compared to *SynD5*. Note that the *p-value* for each data set is calculated only using the accuracies of the model over the test set, which reflects the generalizability of the model.

Comparing simple and proper synthesis

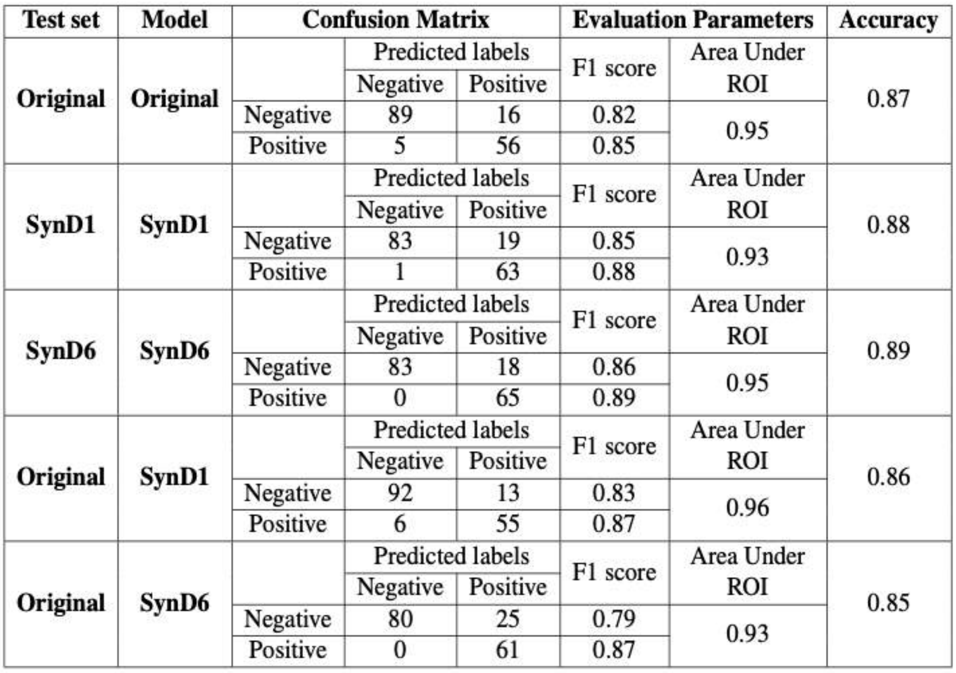
Next, the original data set was synthesized again using *"cart"* method, but this time with setting the proper argument to TRUE *(SynD6)* for posterior predictive distribution of parameters (proper synthesis). Furthermore, a test set of original data sets is fed in the synthetic data fitted model to evaluate the level of local and global structure-preserving capacity of the synthesis method, and pertinence of one aspect of the secondary data analysis. The comparative performance can be seen in Table 5.

The *p-value* from t-test for accuracies of *SynD6* data set in comparison to the accuracies of original data set is *0.0006553*. The *p-value* is smaller than α, which suggests strong evidence to reject the null hypothesis. Furthermore, from the results shown in Table 5, we can use the outcomes to leverage the formal finding. The original data test set performs better for *SynD1* than for *SynD6* data fitted model. Since the difference in F1-score of original test data set is smaller for *SynD1* data fitted model than of *SynD6* data set fitted model.

Visual model comparison

Next, we examined the FI plots for original, *SynD1*, and *SynD6* data sets and ALE plots of first 6 important features for each data set. The number of repetitions was set to 5, defining how often to shuffle features while calculating FI for more stable and accurate results. Figures 22, 23, and 24 show the FI and ALE plots for the original, *SynD1*, and *SynD6* data sets, respectively.

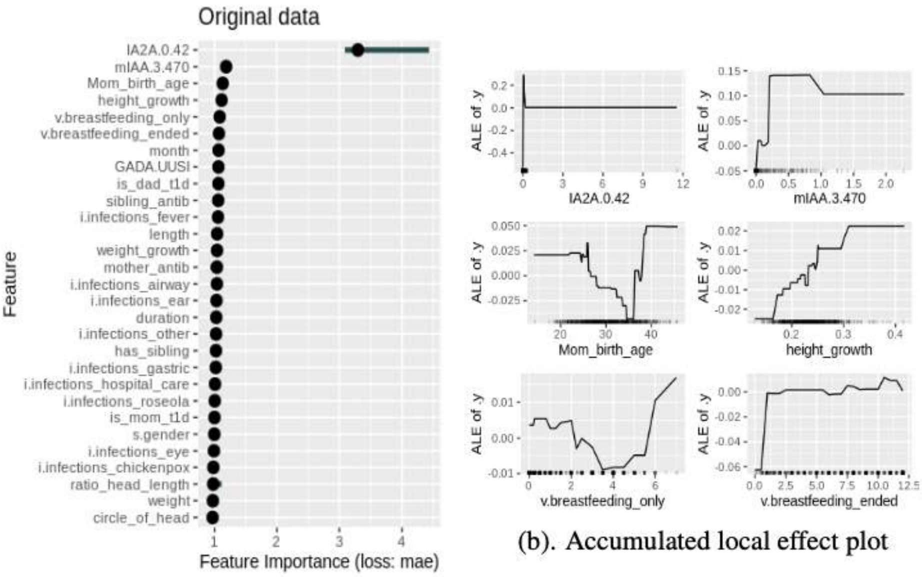
Table 5. Data sets and their performance over GBM models



Test set	Model	Confusion Matrix			Evaluation Parameters		Accuracy
Original	**Original**		Predicted labels		F1 score	Area Under ROI	0.87
			Negative	Positive			
		Negative	89	16	0.82	0.95	
		Positive	5	56	0.85		
SynD1	**SynD1**		Predicted labels		F1 score	Area Under ROI	0.88
			Negative	Positive			
		Negative	83	19	0.85	0.93	
		Positive	1	63	0.88		
SynD6	**SynD6**		Predicted labels		F1 score	Area Under ROI	0.89
			Negative	Positive			
		Negative	83	18	0.86	0.95	
		Positive	0	65	0.89		
Original	**SynD1**		Predicted labels		F1 score	Area Under ROI	0.86
			Negative	Positive			
		Negative	92	13	0.83	0.96	
		Positive	6	55	0.87		
Original	**SynD6**		Predicted labels		F1 score	Area Under ROI	0.85
			Negative	Positive			
		Negative	80	25	0.79	0.93	
		Positive	0	61	0.87		

For original data (Figure 22) FI plot shows that the IA2 antibody values have the most substantial influence in the prediction of the positivity of the autoantibodies later in life, followed by the IAA antibody values, mother's age at the time of birth, height growth rate, age when exclusive breastfeeding ended, and age when any breastfeeding ended. From the ALE plots, we can interpret that after the IA2 value reaches a specific value, the probability of positivity reaches a constant, whereas it decreases with higher IAA value. If the mother's age at the time of birth is higher than approximately 37 years, the probability of positivity increases.

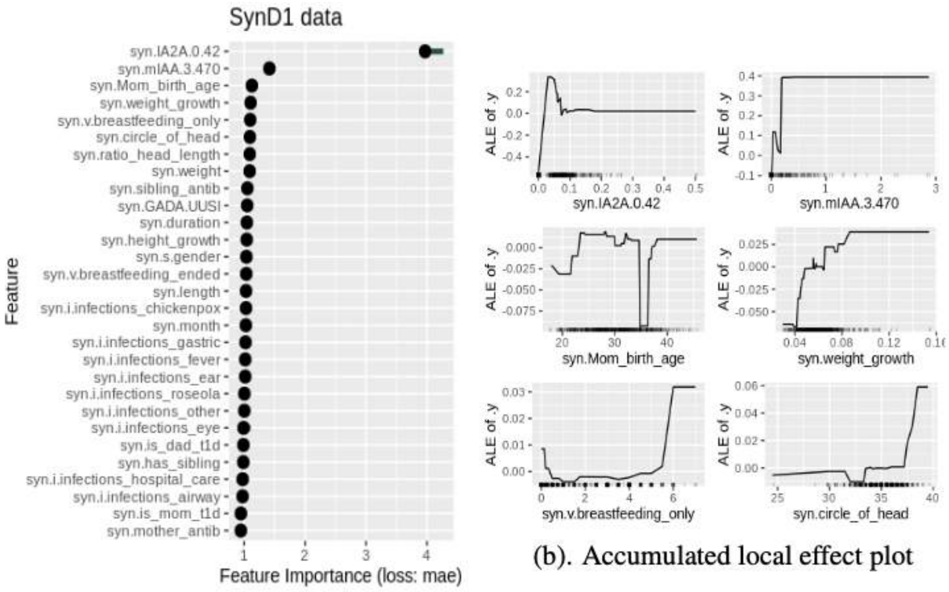
From the FI and ALE plots of original (Figure 23), *SynD1* (Figure 24), and *SynD6* (Figure 25) data set fitted models, we can interpret that *SynD1* data fitted model have higher number of same variables in the first 6 important features and their influence for the model prediction to the original data set fitted model as compared to the *SynD6* data fitted model. As the FI plots for original data set and *SynD1* have same first 3 important features and their ALE plots shows similar accumulated local effect for the matching features including feature *"v.breastfeeding_only"*.



(b). Accumulated local effect plot

(a). Feature importance plot

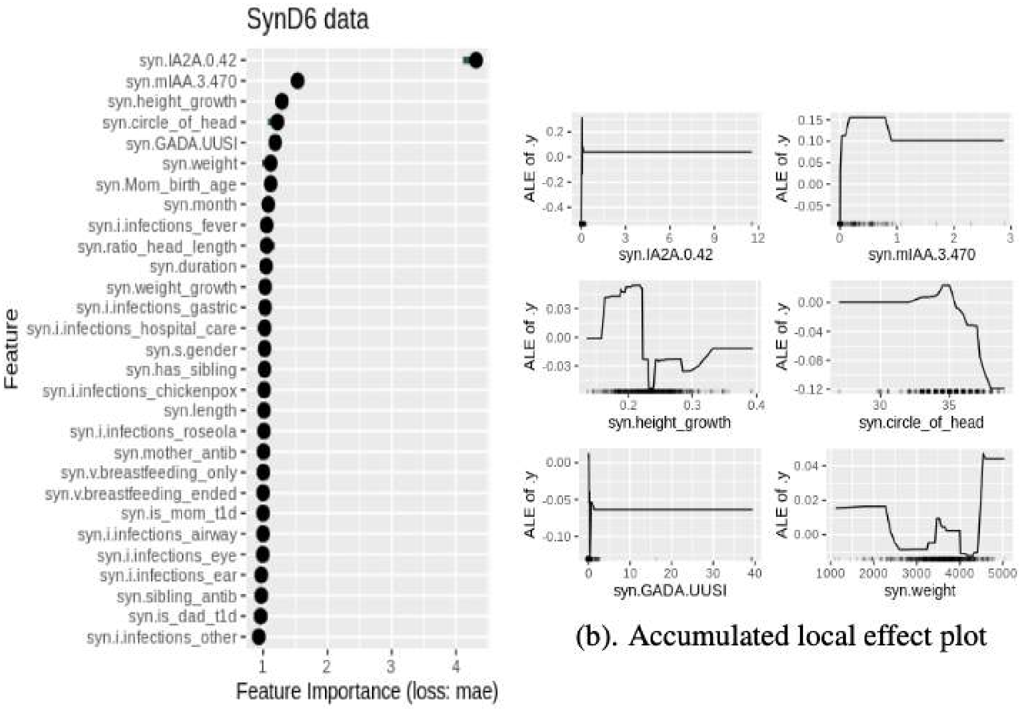
Figure 23. Original DIPP data



(b). Accumulated local effect plot

(a). Feature importance plot

Figure 24. SynD1 data



(b). Accumulated local effect plot

(a). Feature importance plot

Figure 25. SynD6 data

General Utility

Following the performance evaluation of different methods of synthesis based on the specific utility of the original data set, in this section, we analyze and compare the statistical properties of the most reliable synthetic data set (*SynD1*) to statistical properties of the original data set. Comparative analyses include the calculation and measuring the changes in the Pearson correlation between variables, relative frequency distribution, data visualization, and finally, the similarity between original and synthetic data set variables.

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

Pearson Correlation

The PPMCC matrix for original and *SynD1* data set can be seen in Figure 26. The lower triangle represents the Pearson correlation for the original data set, whereas the upper triangle represents the Pearson correlation for *SynD1* data set.

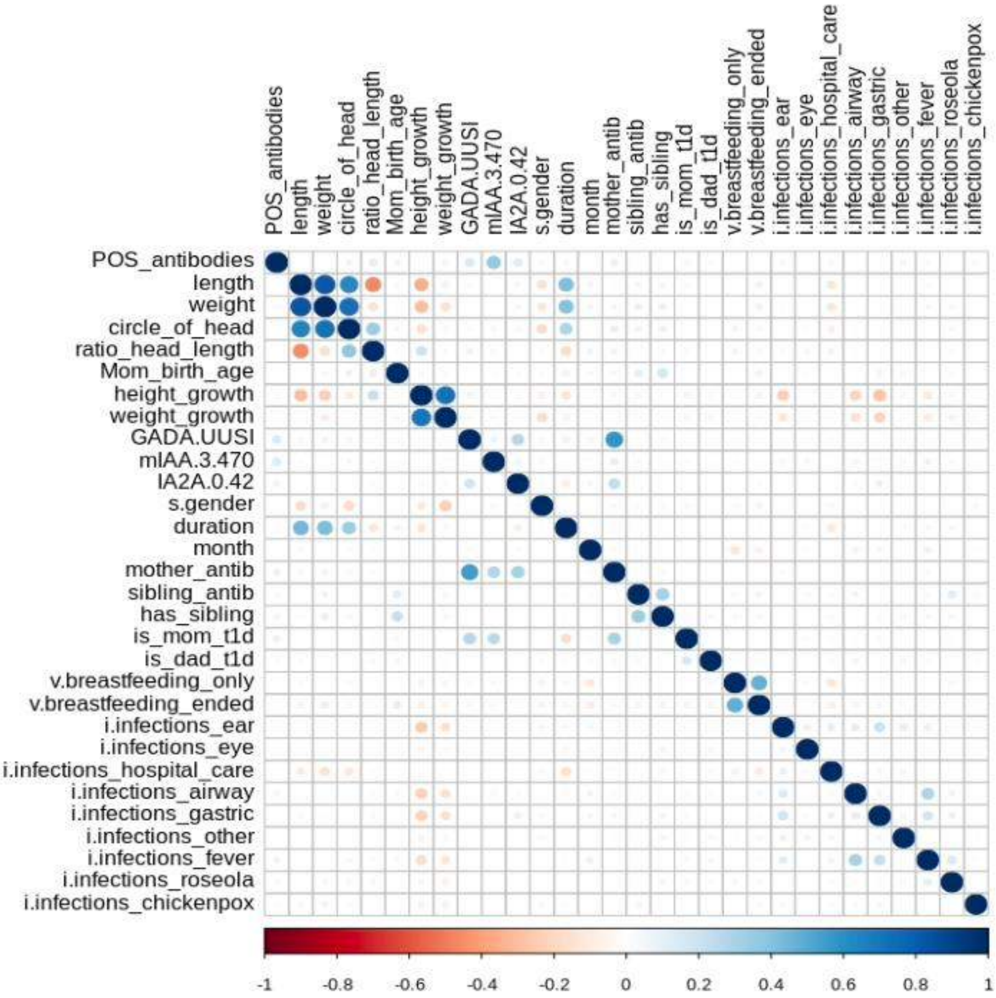


Figure 26. Pearson correlation for original data set in "lower" triangle and SynD1 data set in "upper" triangle

The Pearson correlation in Figure 26 for original and *SynD1* data set looks almost identical. However, the correlation between the variables *POS_antibodies* and IAA antibody are slightly stronger in the *SynD1* data set, *ρ-value* is 0.13 in original and 0.37 in synthetic data set. Furthermore, the correlation between the variables *is_mom_t1d* and *mother_antib* suffered a decrease, with *ρ-value* 0.3 in original and -0.01 in synthetic data set.

Relative frequency distribution

The objective is to evaluate whether and to what degree the data synthesis process preserves the probability distribution of the original data set. The relative frequency distributions of the original data set features in comparison with *SynD1* data set features were plotted to compare the likelihood of a specific result to occur in each population.

Figures 27 shows an example of the relative frequency distribution of a few variables from the original and *SynD1* data sets. The analysis revealed similar distributions between the original and synthetic data sets for every variable.

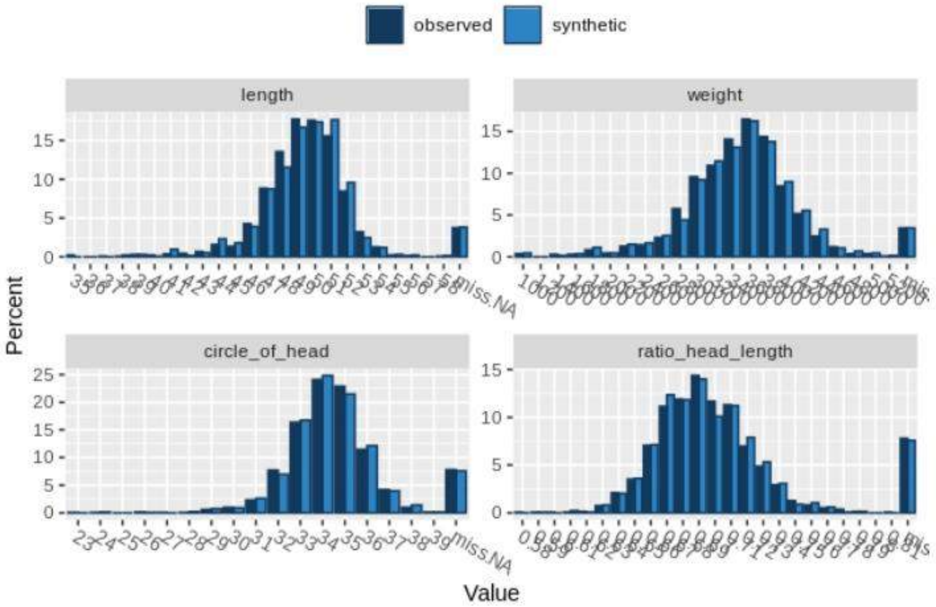


Figure 27. Relative frequency distribution of a few original (observed) and SynD1 (synthetic) data set variables.

Uniform Manifold Approximation and Projection

UMAP can be used for dimension reduction of the data set before fitting the model; however, in our case, we are using UMAP tool to reduce the dimension to be able to visualize the data sets and evaluate the global and local structures within data before and after synthesis.

Figures 28 and 29 shows the UMAP embedding for the original and *SynD1* data set respectively. In both figures, class 1 (positive) samples are represented by red color dots and class 0 (negative) samples with grey color dots. UMAP for the original data set (Figure 28) has approximately four clusters, whereas UMAP for *SynD1* data set (Figure 29) have two distinct clusters with other clusters more scattered as compared to the original data set.

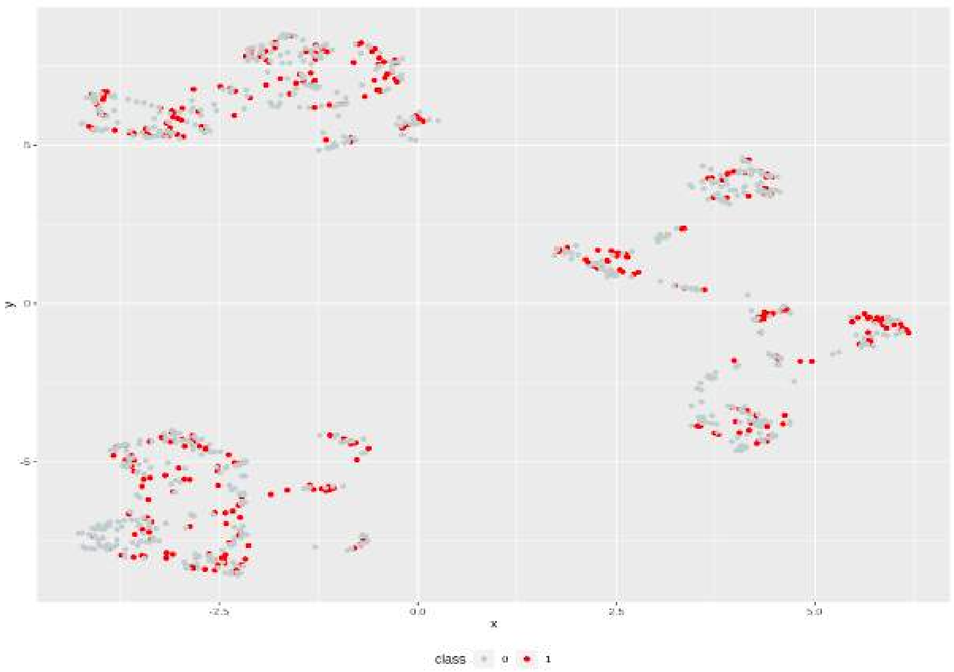


Figure 28. UMAP for original data set

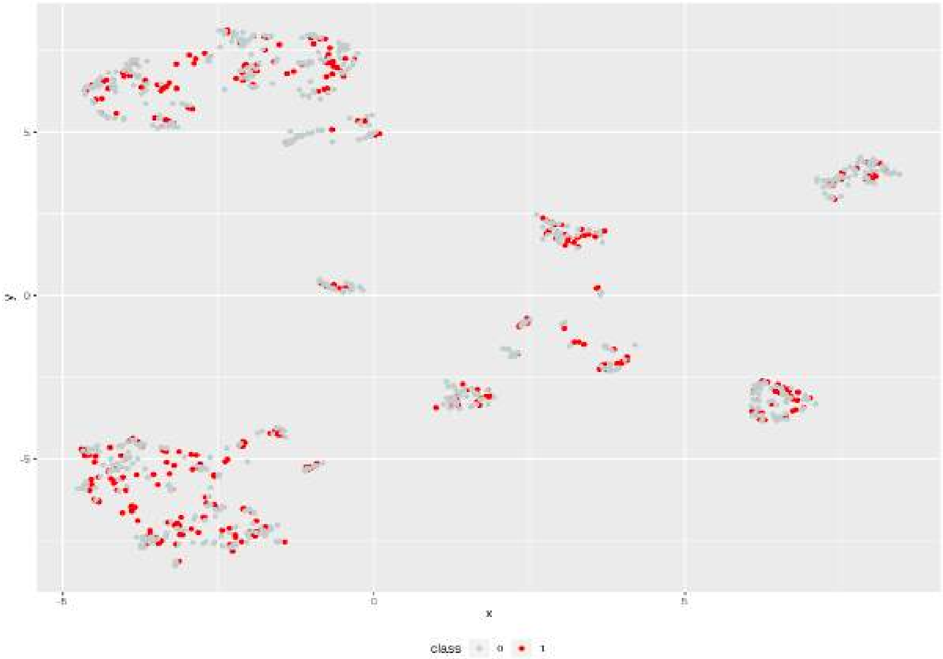


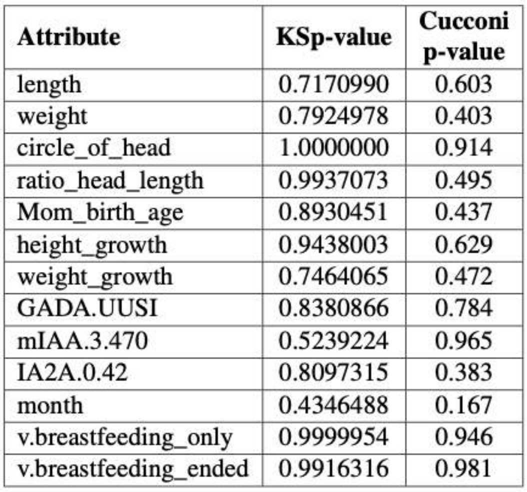
Figure 29. UMAP for SynD1 data set

Data Similarity

Data similarity of the continuous and discrete variables between original and *SynD1* data sets using Kolmogorov–Smirnov two-sample and Cucconi test can be seen in Table 6. From Table 6, all attributes have *kSp-value* and Cucconi *p-value* greater than α, which states that the analysis failed to reject the null hypothesis. In other words, the difference in the distribution of these variables is statistically non-significant.

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

Table 6. KSp-value and Cucconi p-value for matching continuous and discrete attributes between original and SynD1 data sets.



Attribute	KSp-value	Cucconi p-value
length	0.7170990	0.603
weight	0.7924978	0.403
circle_of_head	1.0000000	0.914
ratio_head_length	0.9937073	0.495
Mom_birth_age	0.8930451	0.437
height_growth	0.9438003	0.629
weight_growth	0.7464065	0.472
GADA.UUSI	0.8380866	0.784
mIAA.3.470	0.5239224	0.965
IA2A.0.42	0.8097315	0.383
month	0.4346488	0.167
v.breastfeeding_only	0.9999954	0.946
v.breastfeeding_ended	0.9916316	0.981

Quality of Information Content

After analyzing the impacts of data synthesis and usability of data from a data mining point of view, the concepts of information theory are used further to evaluate the level of distortion in a data set and quantify the information loss.

Entropy

Claude Shannon’s entropy in bits was calculated for each variable in the data set. Figure 29 shows the entropy for each variable in bits before and after synthesis.

© The HTx Consortium 2019-2023. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement Nº 825162.

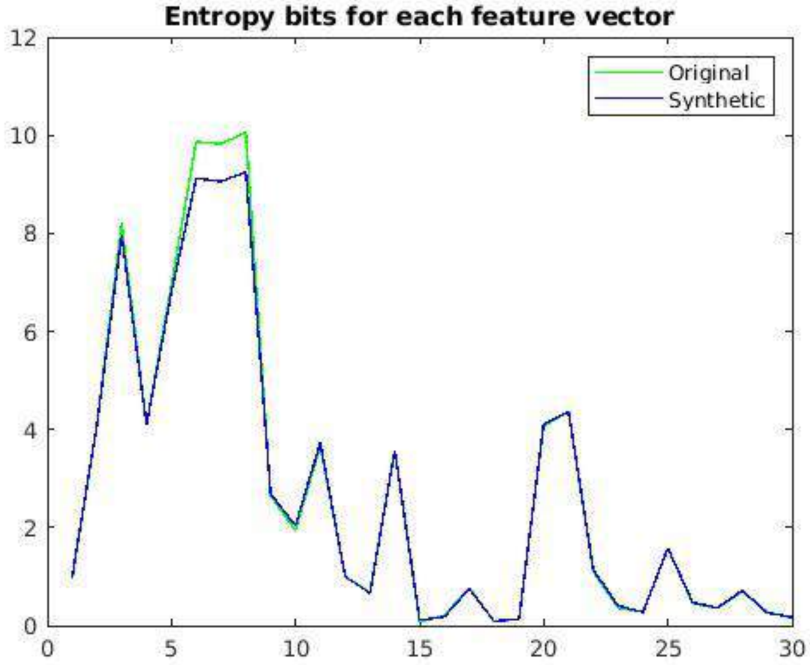


Figure 29. Entropy per bits for original and SynD1 data variables.

From Figure 29, for almost all variables, except a few, entropy remains similar in both original and synthetic data set (*SynD1*). Variables such as the age of mother at the time of birth, growth rate of height and weight had a decrease in entropy by approximately one bit. A decrease in entropy bits states an increase in predictability of the values in the variables.

Mutual Information

Using a variation of kraskov's estimation method, MI for both original and synthetic data (*SynD1*) was calculated between the response variable *POS_antibodies* and rest 29 attributes. The number of neighbors was set to 6 for the k-NN algorithm, and the distance was calculated over a non-Euclidean space. Results revealed that MI remains the same in both data sets.

Outlines

Synthetic data set (*SynD1*) generated using *"cart"* method while not setting the argument proper to TRUE, performs adequately in all the analysis performed. The synthetic data showed no statistically significant difference from the original data set for performance on a machine learning model to predict the positivity of the autoantibodies later in life and dependence on variables. Furthermore, the synthetic data showed a similar bivariate correlation, relative and one-dimensional frequency distribution, and UMAP to the original data set. Additionally, the synthetic data revealed similar entropy bits for each variable and equal mutual information to the response variable.

5.4 Discussion

The objective of the study was to assess the performance of a tool for data synthesis, termed Synthpop by estimating the impacts of the data synthesis process. In order to accomplish the goal, the characteristics of the primary tool of data synthesis were described and utilized to generate synthetic data sets. Next, various data standards were established based on the general and specific utility, and quality of information contained in the original data set. The general utilities are the statistical properties of a data set, whereas the specific utilities are the performances of a data- fitted model. Lastly, the quality of information content is the entropy and MI within a data set. After successfully establishing the data standards, the impacts of the data synthesis process were measured from the differences in the data set before and after synthesis, based on the established standards. A synthetic data set is considered adequate when two requirements are met: 1) the difference or change in the synthetic data utilities as compared to the original data are statistically non-significant and 2) the quality of the information content in synthetic data is similar to that of the original data set.

It is natural to see a healthcare database suffering from imbalanced classes. Especially in real-world data, this is often expected as the data is not gathered in an experimental setting such as randomized controlled trial, instead collected in real-world settings, such as from patient surveys, clinical trials, and observational cohort studies. It is necessary to note that such characteristics affect the performance of machine learning algorithms. In our case, it is not the scope of the study to investigate this reasoning and improve the performance; however, it is essential as it has also affected the data synthesis process. The tool imputes the value for synthetic variable from fitted parameters of synthesizing models, and imbalanced classes played a significant role in most of the synthesizing methods.

The DIPP data set was pre-processed and mostly aggregated from a longitudinal database. When such data is generated, it is expected that the data set will suffer from imbalanced classes. Such characteristic plays a significant role in data analysis; in our case, the effects can be seen in model training. Most of the synthetic data-fitted models, including the original data-fitted model performed reasonably well in the prediction of negative and positive cases despite having a high number of negative samples. However, one synthetic data set (*SynD5*) significantly favored the negative samples more than any other data set. The *SynD5* was the only synthetic data set which was produced using parametric methods fitting the type of data variables. This analysis suggests

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

that the during data synthesis, model fitting parameters of the synthesizing method might have suffered overfitting, and synthetic data values were imputed to favor negative classes. Even though the significance test of the accuracies of the SynD5 data-fitted model reports no statistically significant difference to the original data-fitted model, when other evaluation parameters were considered, the *SynD5* revealed various shortcomings favoring the previous finding. These interpretations underline the importance of the other evaluation parameters while determining a model's performance.

On the other hand, non-parametric methods have generated synthetic data which not only fell behind but also exceeded the performance of the original data-fitted model. For example, *SynD2*, *SynD3* and *SynD4* synthetic data-fitted model show the accuracies of 86.0%, 90.0% and 93.0%, respectively. The difference in the performance could be again due to overfitting or underfitting of model fitting parameters of synthesizing methods. However, it could also be induced by a variation in bivariate correlations between variables during data synthesis. All these data sets (*SynD2*, *SynD3*, and *SynD4*) also showed a statistically significant difference in model accuracies with the original data set. From these analyses, we can say that the specific utility of synthetic data is highly dependent on the method of synthesis. The issues could be resolved by thoughtfully selecting a different method for data synthesis according to the type of data. Moreover, we can also control the model fitting parameters of the synthesizing method by controlling the *predictor.matrix* to avoid overfitting and underfitting of the synthesizing model.

Despite its weaknesses, the tool exceeded the expectations when the default method of synthesis *"cart"* was used, which can handle any data type. Two synthetic data sets were generated using *"cart"* method: *SynD1* and *SynD6*. The only difference was that *SynD6* data was generated while setting the argument proper to TRUE for proper synthesis. Repeatedly, the *SynD6* data-fitted model showed signs of overfitted parameters of synthesizing model during data synthesis—however, the *SynD1* data- fitted model outperformed in all analyses. The synthetic data set showed no signs of variation in data utility. Synthetic data set *SynD1* succeeded at all performed tests with the statistically non-significant difference from the original data set; this is the only synthetic data set which leads to failing to reject the null hypothesis. Additionally, the quality of the information content was also well preserved for 27 out of 30 variables. For the rest of the three variables, *SynD1* suffered a decrease in entropy only by 1-bit. Conclusively, these analyses suggest that the "cart" method not only preserved the utilities but also preserved the complexity of the DIPP data set according to the data standard established in this study; exhibiting that the tool certainly accomplished its intended goal.

5.5 Synopsis and Future Work

The impediments in healthcare data mining and sharing most often relate to research participant's or patient's privacy, security, and the circumstance that researchers face of having to consider the trade-off between the risk of disclosure and the benefits of open data sets [7, 26, 27, 28]. Open healthcare data not only benefits extended scientific collaboration for innovative discoveries and validating previously defined hypotheses but more importantly, sharing healthcare data could save lives. Healthcare databases are demonstrating to play an indispensable part in controlling and preventing the spread of the novel coronavirus "COVID-19" (SARS-CoV-2) in a worldwide pandemic [29, 30, 31, 32]. In numerous situations, the survival of a database itself depends on the data holder's capability to provide data when needed, since not releasing such data at all may eventually diminish the need for it [8]. However, the process of sharing healthcare data needs careful measures as it could unfold severe consequences through the risk of disclosure and could harm not only the participants but also organizations or individuals involved in collecting and sharing data [33].

Current data sharing systems, including SQLShare [34] and DataHub [35], promote collaborative data analyses but fail to consolidate privacy-preserving prospects or means to manage sensitive data. Synthpop could amend this by producing a synthetic version of the original data set. Furthermore, the use of synthetic data for secondary data analysis will enhance the collaboration between data owners and external data scientists while maintaining the subject's privacy. However, the achievement of anonymity relies on the assumption that there are no matching samples between the original and synthetic data sets, also, there are no samples with extreme values which could serve as a unique identifier. Additionally, the utility of the data is highly dependent on the performance of the synthesizing model, and the Synthpop package itself provides minimal tests for the synthetic data analysis. Comparing only the relative frequency distribution of two data sets for statistical analysis says a lot but from a rather vague perspective. Furthermore, the package provides the comparison of the data-fitted models but only with linear machine learning techniques which is again somewhat limited.

However, as demonstrated in this study, a user could utilize different tools to measure the utility of the data or consolidate further questioning if desired. Subsequently studying and assessing Synthpop by measuring the impacts of the data synthesis process, we conclude that the tool performs competently in the current setting. Future researchers could consider implementing a more sophisticated way to read entropy bits and investigating the mutual information between pairs of variables in both original and synthetic data sets could highlight more in-depth impacts of the data synthesis process. Further analyses, including the involvement of different tools for data anonymization such as differential privacy, can provide a comparative analysis from a different point of view. Subsequently, examining the performance of Synthpop on longitudinal data could provide a greater understanding of the comprehensiveness of the tool. Finally, more advanced analyses are required to question whether the assumed anonymity in the synthetic data set exists, and to what extent.

5.6 Conclusion

The study was inspired by the benefits of open healthcare databases and aimed to examine a unique solution to perpetual hindrances in data sharing caused by the risk of disclosure and shortcomings of current data anonymization techniques. Therefore, in this study, the performance of a tool for data synthesis, termed Synthpop, was analyzed by assessing the impacts of the data synthesis process. Impacts were measured based on the quantifiable changes in utilities and quality of information contained in the data set before and after synthesis. Two different types of data sets were used to evaluate the generalizability of the data synthesis tool. Our statistical analyses conclude that the tool is generalized in terms of applicability to different data types. Furthermore, synthetic data mimics the original data set while preserving all the statistical properties, machine learning capabilities, and quality of the information contained in the original data set with statistically non-significant differences. In conclusion, the tool succeeded at its intended purpose and can be used to generate synthetic data sets for data sharing purposes. However, the performance of the tool profoundly depends on the method of synthesis, i.e., carefully choosing a method of synthesis improves the performance of the tool.

Overall, Synthpop fulfils all the necessities towards data sharing and hence unfolds a wide range of opportunities in the research community, including easy data sharing, more significant collaborations, and information protection [26]. Considering the workflow of the study, we can also state that data collectors and authors will always be indulged, since the findings from the synthetic data need verification from the original data set. This dependency on the original data set for result verification embeds a limitation on the study because the synthetic data can only be used for secondary data analysis. If the original author cannot be reached for result verification, the analyses may cease and result in an abandoned study.

6. Conclusion and future work

This deliverable concentrated on gaining knowledge from medical data via data exploration and visualization. It was shown how visualization makes the understanding of data structure a lot easier and this understanding can lead to finding associations and explanatory factors from the data which then can be used to model treatment effectiveness. This deliverable presented ClinFlow, a visualization tool containing an easy access to several methods to process medical data and unsupervised machine learning methods to illustrate different aspects of data. In addition, this deliverable presented how understanding of data structure can be used to create anonymized synthetic data. The main advantage of synthetic data is that it is anonymized, due to this, it provides an opportunity to share data between partners without risk of privacy issues and GDPR violations.

Data-driven models are extremely good at discovering associations in the data. Unfortunately, they cannot guarantee causality of these associations. Providing acceptable explanations of the reasons behind the model's behavior as well as the opportunity for the expert to intervene will contribute to the trustworthiness and safety of the methods and will improve the understanding behind the effects of choosing between different treatment alternatives. Ultimately, explainable AI will bring us closer to employing these technologies for real life decision support in medicine. In fact, explainable AI will be studied in upcoming WP3 deliverables. Moreover, the objective of future work is to provide accessible tools for prediction of individual patient treatment outcomes, personalized treatment pathways, disease risk estimation and identification of risk groups, aimed at the medical experts. These tools will implement transparent and interpretable ML and AI models into user interfaces with interactive visualizations of the explanations behind model behaviors, and options to allow users to interact with the models for improved predictions.

Acknowledgement

DIPP data is collected and owned by University of Oulu, Ostrobothnia Hospital district, University of Turku, Varsinais-Suomi Hospital District, University of Tampere, and Pirkanmaa Hospital District. The data gathering is organized by the PIs Riitta Veijola, Jorma Toppari, Jorma Ilonen, Johanna Lempainen, Mikael Knip, Kalle Kurppa, and Heikki Hyöty.

References

[1] Wikipedia: Dimensionality reduction.
https://en.wikipedia.org/wiki/Dimensionality_reduction, Accessed 16.04.2021.

[2] Finnish diabetes association. URL: https://www.diabetes.fi, Accessed 21.01.2020.

[3] Haller, MJ, Schatz, DA. The DIPP project: 20 years of discovery in type 1 diabetes. Pediatr Diabetes,2016;17: 5-7.

[4] Knip M, Siljander H, Ilonen J, Simell O and Veijola R. Role of humoral beta-cell autoimmunity in type1 diabetes. Pediatr Diabetes, 2016;17: 17-24.

[5] Pöllänen PM, Lempainen J, Laine AP, Toppari J, Veijola R, V˙ah¨asalo P, Ilonen J, Siljander H, Knip M.Characterisation of rapid progressors to type 1 diabetes among children with HLA-conferred diseasesusceptibility. Diabetologia. 2017;60(7):1284-1293.

[6] Van Ginneken A.M. (2002) The computerized patient record: balancing effort and benefit. International Journal of Medical Informatics 65, pp. 97–119.

[7] Cios K.J. & Moore G.W. (2002) Uniqueness of medical data mining. Artificial Intelligence in Medicine 26, pp. 1–24.

[8] Sweeney L. (2002) k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, pp. 557– 570. DOI: 10.1142/S0218488502001648.

[9] Viceconti M., Hunter P. & Hose R. (2015) Big data, big knowledge: big data for personalized healthcare. IEEE Journal of Biomedical and Health Informatics 19, pp. 1209–1215. DOI: 10.1109/JBHI.2015.2406883.

[10] Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA l. Rev., 57, 1701.

[11] Anguita D., Ghio A., Oneto L., Parra X. & Reyes-Ortiz J.L. (2013) A public domain dataset for human activity recognition using smartphones. In: Proc. 2013 ESANN, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges (Belgium), pp. 437–442.

[12] Nowok B., Raab G.M., Dibben C. et al. (2016) Synthpop: Bespoke creation of synthetic data in r. Journal of Statistical Software 74, pp. 1–26. DOI: 10.18637/jss.v074.i11.

[13] Snoke J., Raab G., Nowok B., Dibben C. & Slavkovic A. (2018) General and specific utility measures for synthetic data. Journal of the Royal Statistical Society. Series A: Statistics in Society 181, pp. 663–688. DOI: 10.1111/rssa.12358.

[14] Maaten L.v.d. & Hinton G. (2008) Visualizing data using t-sne. Journal of Machine Learning Research 9, pp. 2579–2605.

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

[15] McInnes L., Healy J. & Melville J. (2018) Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426v2.

[16] Schapire R.E. (1990) The strength of weak learnability. Machine Learning 5, pp. 197–227. DOI: 10.1007/BF00116037.

[17] Freund Y. (1995) Boosting a weak learning algorithm by majority. Information and Computation 121, pp. 256–285. DOI: 10.1006/inco.1995.1136

[18] Freund Y., Schapire R.E. et al. (1996) Experiments with a new boosting algorithm. In: 13th International Conference proceedings, Machine Learning, pp. 148–156.

[19] Matlab (R2019b) Deep Learning Toolbox. The MathWorks, Inc., Natick, Massachusetts, United State. URL: https://se.mathworks.com/ products/statistics.html, Accessed 02.02.20.

[20] Matlab (R2019b) Statistics and Machine Learning Toolbox. The MathWorks, Inc., Natick, Massachusetts, United State. URL: https://se.mathworks. com/products/deep-learning.html, Accessed 02.02.20.

[21] Fisher R.A. (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, pp. 179–188. DOI: 10.1111/j.1469-1809.1936.tb02137.x.

[22] Oliver D.I. (2014) Privacy engineering: A dataflow and ontological approach. CreateSpace Independent Publishing Platform.

[23] Oliver I. & Miche Y. (2016) On the development of a metric for quality of information content over anonymised data-sets. In: Proc. 2016 IEEE, 10th International Conference on the Quality of Information and Communications Technology (QUATIC), pp. 185–190. DOI: 10.1109/QUATIC.2016.047.

[24] Kraskov A., Stögbauer H. & Grassberger P. (2004) Estimating mutual information. Physical Review E 69, p. 066138. DOI: 10.1103/PhysRevE.69.066138.

[25] Shannon C.E. (1948) A mathematical theory of communication. Bell System Technical Journal 27, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[26] Quintana D. (2019) Synthetic datasets: A non-technical primer for the behavioural sciences to promote reproducibility and hypothesis-generation DOI: 10.31234/osf.io/dmfb3.

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.

[27] Lenert L. & McSwain B.Y. (2020) Balancing health privacy, health information exchange and research in the context of the covid-19 pandemic. Journal of the American Medical Informatics Association

[28] Ienca M. & Vayena E. (2020) On the responsible use of digital data to tackle the covid-19 pandemic. Nature Medicine 26, pp. 463–464.

[29] United Nations, Department of Economic and Social Affairs, News (2019). COVID-19 – when data save lives. URL: https://www.un.org/ development/desa/en/news/statistics/covid-19-when- data-save-lives.html, Accessed 15.05.2020.

[30] Gates B. (2020) Responding to covid-19—a once-in-a-century pandemic? New England Journal of Medicine 382, pp. 1677–1679.

[31] Kucharski A. J., Russell T.W., Diamond C., Liu Y., Edmunds J., Funk S., Eggo R.M., Sun F., Jit M., Munday J.D. et al. (2020) Early dynamics of transmission and control of covid-19: a mathematical modelling study. The Lancet Infectious Diseases.

[32] Wu Z. & McGoogan J.M. (2020) Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china: summary of a report of 72 314 cases from the chinese center for disease control and prevention. The Journal of the American Medical Association 323, pp. 1239–1242.

[33] Baker & McKenzie (2020) COVID-19 Data Privacy & Security Survey. URL: https://www.bakermckenzie.com/-/media/files/insight/publications/2020/04/covid19-data-privacy--security- survey17-april.pdf, Accessed 20.05.20.

[34] Jain S., Moritz D., Halperin D., Howe B. & Lazowska E. (2016) Sqlshare: Results from a multi-year sql-as-a-service experiment. In: Proc. 2016 ACM, International Conference on Management of Data, pp. 281–293.

[35] Bhardwaj A., Bhattacherjee S., Chavan A., Deshpande A., Elmore A. J., Madden S. & Parameswaran A.G. (2014) Datahub: Collaborative data science & dataset version management at scale. arXiv preprint arXiv:1409.0798.

[36] Chen D, Fu LY, Hu D, Klukas C, Chen M, Kaufmann K. TheHTPmod Shiny application enables mod-eling and visualization of large-scalebiological data. Communications Biology 2018;1.

[37] Kahn H.S. & Morgan T. (2009) Association of Type 1 Diabetes With Month of Birth Among U.S. Youth. Diabetes care 32(11): 2010-5. DOI: 10.2337/dc09-0891.

[38] Virtanen S., Takkinen H.M., Nwaru B., Kaila M., Ahonen S., Nevalainen J., Niinistö S., Siljander H., Simell O., Ilonen J., Hyöty H., Veijola R. & Knip M. (2014) Microbial Exposure in Infancy and Subsequent Appearance of Type 1 Diabetes Mellitus–Associated Autoantibodies. JAMA Pediatrics 168(8): 755-763. DOI: 10.1001/jamapediatrics.2014.296

[39] Cardwell C., Stene L., Joner G., Bulsara M., Cinek O., Rosenbauer J., Ludvigsson J., Jané M., Svensson J., Goldacre M., Waldhör T., Jarosz-Chobot P., Gimeno S., Chuang L.M., Parslow R., Wadsworth E., Chetwynd A., Pozzilli P., Brigis G. & Patterson C. (2010) Maternal Age at Birth and Childhood Type 1 Diabetes: A Pooled Analysis of 30 Observational Studies. British Journal of Nutrition 59: 486-494. DOI: 10.2337/db09-1166.

[40] RStudio, Inc (2013) Easy web applications in R. URL: http://www. rstudio.com/shiny/. Accessed 01.02.2020.

© The HTx Consortium 2019-2023. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825162.