ExploreR Dataset Structure Guide

Thank you for your interest in the ExploreR. This documents aims to explain the structure of the data tables used by the ExploreR. If available, we recommend the synthetic data included with the tool as an example.

The ExploreR uses linked data. This consists of two tables: Attributes and Activity. Both are required. Please note, in the event all the data outlined in this guide is not available, it is possible to use dummy data / reduced quantities to run the ExploreR regardless.

**Attributes Table**

This section covers the structure of the **Attributes** **table**.

This table contains a record for each individual patient in the system, where a row gives all the attributes of the individual, and other rows relate only to other patients. The following data is required (at minimum):

1. Sex (please note “male” and “female” must be both present in this field, and be in lowercase as in the quotation marks here)
2. Age
3. Ethnicity
4. Deprivation (an overall measure, i.e. Index of Multiple Deprivation)
5. Some geographical identifier (e.g. locality/ICP)
6. Some clinical conditions

[Bridges to Health (BtH](https://outcomesbasedhealthcare.com/bridges-to-health-segmentation-model/)) segmentation is integrated using the following definitions:

|  |  |
| --- | --- |
| Segment | Definition |
| Frailty | Is in Electronic Frailty Index (EFI) category 'Frail' or 'Moderate' |
| Limited Reserve | Has one or more of: chronic kidney disease, heart failure, or ever had a myocardial infarction |
| Short Period of Decline | Has had any form of cancer at some time since 2003 |
| Stable But Serious Disability | Has one or more of: hearing impairment, visual impairment, ataxia, amnesia, aphasia, cerebral palsy, or a brain injury |
| Chronic Conditions | Has one or more of: ischaematic heart disease (excluding myocardial infarction), arrhythmia (excluding atrial fibrillation), hypertension, diabetes (type 1 or 2), non-alcoholic fatty liver disease, depression, serious mental illness, personality disorder, inflammatory arthritis, asthma, chronic obstuctive pulmonary disorder, or an 'other' significant cardiovascular condition (not explicitly described) |
| Maternal | Is female and pregnant |
| Acutely Ill | Has attended A&E (for any reason) within the previous 12 months |
| Healthy | Anyone not covered by the segments above |

If the user has a different definition in mind, then simply pre-calculate the definitions, and for each segment select the corresponding value. Please note the user will need to inform other users the definitions used are different from the contents of the BtH page sidebar. For ease of use, it is recommended that BtH segments are calculated before the data is uploaded to the ExploreR, as this makes segment definitions far easier (the precalculated BtH field can be used to very quickly define the BtH segments in the program).

A&E data should be included in the Activity table.

Almost every field name in the table is prefixed by one of 4 values, designed to tell the ExploreR what it contains: “demog.”, “clinic.”, “area.”, “socio.”. Please note ‘.’ is a reserved character, used to separate prefixes and column names.

1. With the exception of ‘id’, every field name must be prefixed by one of ‘demog’, ‘clinic’, ‘clinic.misc’, ‘area’ or ‘socio’. Fields not prefixed will be ignored.
   1. The ‘id’ field should contain numeric ids for uniquely identifying each patient. This field should match the IDs to those in the ‘id’ field of the activity table
   2. Fields prefixed by ‘clinic.misc’ are aggregated differently from other “clinic.” fields. In addition, ‘clinic.misc’ fields are not required to be binary, and may take any value.
2. ‘demog.’ columns should contain demographic data, including age and sex.
3. ‘clinic.’ columns should contain clinical data. With the exception of ‘clinic.misc’ fields, clinical conditions should be binary, 1 or 0 (where 1 means the individual has the condition). A subset of these values is aggregated to produce multimorbidity scores. Continuous clinical values, for example polypharmacy, should be prefixed with ‘clinic.misc’.
4. ‘area.’ columns should contain geographical/area data, e.g. locality, practice, ICS
5. ‘socio.’ columns should contain socio-economic data, such as deprivation deciles. Please note, discrete fields are required for deprivation measures ( as opposed to continuous scores), as plotting continuous values can be challenging at times.

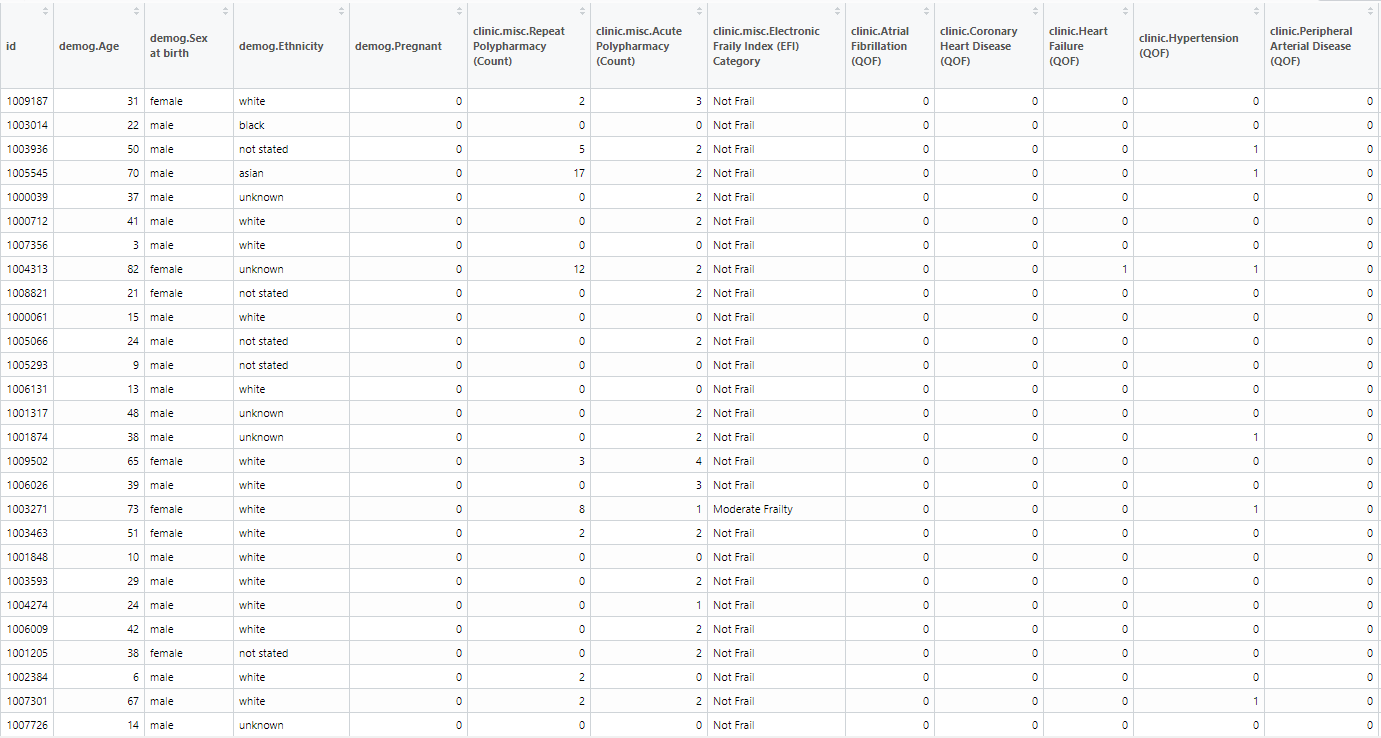


Figure 1 Synthetic data Attributes table

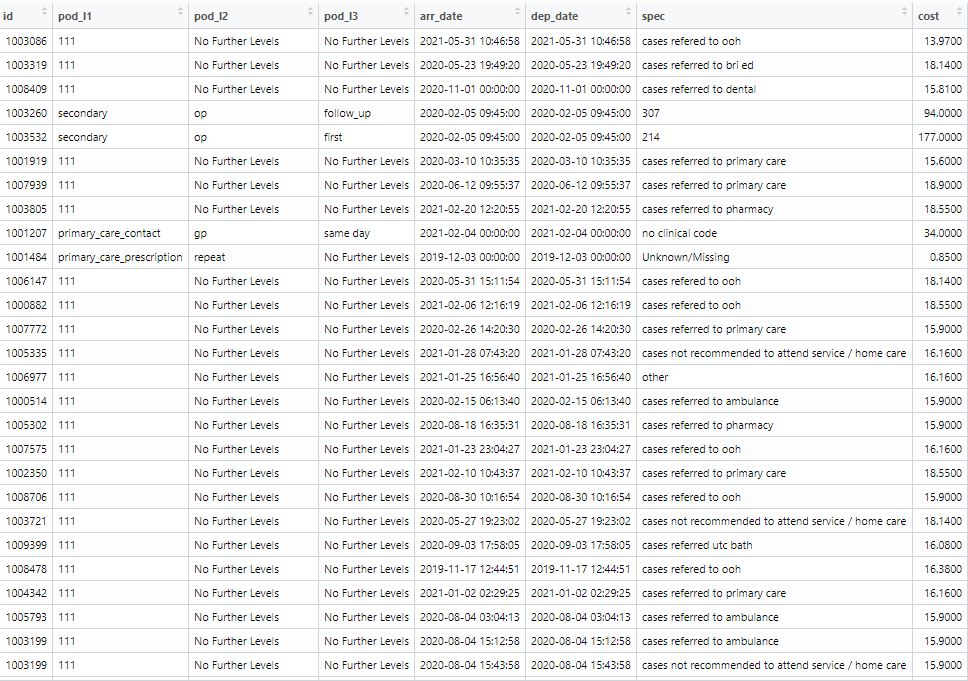


Figure 2 Synthetic data Activity table

**Activity Table**

This section covers the structure of the **Activity** **table**.

In the Activity table, each record represents a point of activity, such as being admitted to A&E, visiting a GP, etc. Unlike the **Attributes Table**, the fields in this table, and their names, are mostly set. Thus, the exact column names present in the *Column name* field in the table below must be present in the Activity table when uploaded to the ExploreR.

|  |  |  |
| --- | --- | --- |
| Column name | Description | Values |
| id | A unique identifier for each patient (not record). Should match with the id field in Attributes Table | Integer |
| pod\_l1 | Descriptor for Point of Delivery (POD). String to classify type of activity (e.g. “primary care”) | String |
| pod\_l\* | Further levels, replace \* by various integers are required. Further descriptor to previous POD(s) | String |
| arr\_date | Arrival Date. Format: see below. YYYY-MM-DD format. Should read as type POSIXct in R. | Date |
| dep\_date | Departure date. If no data, set to be the same as arr\_date Format: see below. | Date |
| Cost | Cost of activity | Numeric |
| Spec | A description of specialisation of the service | String |

The purpose of pod\_l\* are to allow the user to classify activity into groups, so that in the ExploreR particular types of activity may be highlighted.

arr\_date /dep\_date format: If using .**rds** extension files, data type should be POSIXct in format YYYY-MM-DD HH:MM:SS (use ‘00’ for missing HH:MM:SS values). If using .**csv** extension files, data should be in format DD/MM/YYYY HH:MM.

**Package and R Versions**

The ExploreR was built using the following package versions. Please note later versions should, in most cases, be backwards compatible, but this cannot be guaranteed. Packages versions were obtained using *getNamespaceVersion(libraryName)* in R.

R version 3.6.3, on Windows 10

RStudio version 1.3.1093

shiny: 1.6.0

shinydashboard: 0.7.1

tidyverse: 1.3.0

lubridate: 1.7.9.2

shinyBS: 0.61

plotly: 4.9.3

treemap: 2.4-2

RColorBrewer: 1.1-2

shinyTree: 0.2.7

ggplot2: 3.3.5

shinyWidgets: 0.6.0

dplyr: 1.0.6

scales: 1.1.1

rpart: 4.1-15

rpart.plot: 3.0.9

cluster: 2.1.0

FactoMineR: 2.4

data.table: 1.13.2

DT: 0.16