# The PHM ExploreR: packages and raw data information sheet

### Authors: Andras Varady and Anna Powell (BI Modelling and Analytics)

## Introduction

This document details the requirements for the data tables used by the PHM ExploreR. We have also provided a synthetic dataset which can be used as a reference template.

ExploreR uses linked data consisting of two data files: **Attributes** and **Activity**. Unique patient identifiers (e.g., Pseudo-NHS numbers) link the two files. Both files are required for the tool to work. Please note, if your dataset does not include all the required fields, it is possible to use dummy variables in place of the missing variables. This will be explained in more detail later in this document.

## Required software

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 | shinyBS: 0.61 | scales: 1.1.1 |
| RStudio version 1.3.1093 | dplyr: 1.0.6 | rpart: 4.1-15 |
| shiny: 1.6.0 | plotly: 4.9.3 | rpart.plot: 3.0.9 |
| shinydashboard: 0.7.1 | treemap: 2.4-2 | cluster: 2.1.0 |
| tidyverse: 1.3.0 | RColorBrewer: 1.1-2 | FactoMineR: 2.4 |
| shinyTree: 0.2.7 | ggplot2: 3.3.5 | data.table: 1.13.2 |
| shinyWidgets: 0.6.0 | lubridate: 1.7.9.2 | DT: 0.16 |
| VIM: 6.1.1 | MASS: 7.3-55 |  |

## Attributes table

This table contains a record for each individual patient in the system, where a row gives all the attributes of an individual. The following data fields are required (at minimum):

1. ID (e.g. NHS number)
2. Sex (please note “male” and “female” must be both present in this field, and be in lowercase as in the quotation marks here)
3. Age
4. Ethnicity
5. Deprivation (e.g., Index of Multiple Deprivation)
6. At least one geographical identifier (e.g., locality/ICP)
7. At least one clinical condition

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. Each field name in the table, except for ‘**id’**, must be prefixed by one of four values specifying the type of data contained in that field. These four prefixes are: ‘**demog**.’ (demographic data, e.g., age, sex etc), ‘**clinic**.’ (clinical conditions), ‘**area**.’ (geographical region, e.g., ICS, locality, etc), ‘**socio**.’ (socio-economic data, e.g., deprivation deciles[[1]](#footnote-1)). Fields not prefixed with one of these definitions will be ignored. Please note ‘.’ is a reserved character used to separate prefixes from specific column names.

‘**clinic**.’ columns should contain clinical data. With the exception of ‘**clinic.misc**’ fields, clinical conditions must be binary, 1 or 0 (where 1 means the individual has the condition). A subset of these values is aggregated to produce multimorbidity scores. Fields prefixed by ‘**clinic.misc**’ are aggregated differently from other “**clinic**.” and may take any numerical value. Therefore continuous clinical values, for example polypharmacy, should be prefixed with ‘**clinic.misc**’.

### Population segmentation

Bridges to Health (BtH) segmentation is integrated into the tool using the definitions summarised in the table below.

|  |  |
| --- | --- |
| 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 |

For more information on Bridges to Health please see [here](https://outcomesbasedhealthcare.com/bridges-to-health-segmentation-model/)

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).

## 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 PHM 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. 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 YYYY-MM-DD format. Should read as type POSIXct in R. | Date |
| Cost | Cost of activity | Numeric |
| Spec | A description of specialisation of the service | String |

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

## Appendix 1: Screen shot of synthetic attributes table data

Table

Description automatically generated

## Appendix 2: Screen shot of synthetic activity table data

Table

Description automatically generated

1. Please note, discrete fields are required for deprivation measures (as opposed to continuous scores), as plotting continuous values can be challenging at times. [↑](#footnote-ref-1)