## Data Pre-processing

### 1. Data cleaning

Data cleaning is an essential method of improving data quality and reducing the error rate of data in the statistical process. For CleverCogs data, we filtered them in the following three steps and ended up with a table with 540837 rows and 9 columns of activity data, a table with 2664 rows and 6 columns of user information data, and a table with 1310 rows and 8 columns of medical condition data.

#### 1) Null data

Since the ExternalID in each table is the primary key that identifies the user to the connected tables, a null value does not meet the specification. We removed all data in the three tables where the ExternalID was null.

#### 2) Negative data

For some data items with constant positive values, such as ExternalID and Age, we filtered and deleted the parts that are less than or equal to zero in three tables.

#### 3) Data types

For users’ age, we know from the platform data specifications that CleverCogs is only for adults over the age of 18. Therefore we removed all data on minors. For users’ gender, we noticed that it is divided into three categories, female, male, and unspecified. We were unsure during the data processing whether unspecified gender referred to the non-binary or users preferred not to say. We assume that these users with unspecified gender are all machines rather than human beings. As you know, the gender category is an important factor when analyzing activities and medical conditions. To ensure the accuracy of the data analysis results, we have removed user information that does not identify a defined gender. The same applies to users who lack a date of birth.

### 2. Data classification

We consider every activity under the category with the Internet and Category that has not been correctly classified. Also, because the volume gap between categories is so large, it may not be possible to present a clear picture of the results of the data analysis. Therefore, we have broken down the activity data into an additional 41 categories based on the characteristic of each activity.