Minutes of meeting between Catherine Pulman from Chase Africa (CP) and Maja Založnik of OIPA (MZ) in London 17.12.2018.

We discussed options for knowledge exchange both in terms of the data that is currently available for analysis as well as in terms of future data collection. MZ suggested several rules of good practice and processes that should be put into place with the current data pipeline, that will provide a good base for future expansions of the data collection. These are (in no particular order):

The Data Pipeline:

I use the term data pipeline broadly to refer to the whole process of how the data are collected, entered, transferred, stored, processed, analysed, visualised, disseminated.. Thinking explicitly about the data pipeline means separating conceptually the processes of data input, data storage, data processing and data analysis and thinking through the requirements and tools necessary for each step.

Data Management Plan:

Chase Africa should prepare a DMP that formalises some of the aspects of the data pipeline mentioned above, but more importantly the principles that support a good data pipeline and are flexible and adaptable to future needs. CP will look into similar organisations' set-ups for inspiration, but at a minimum such a document should set out commitments to the safe storage of the data, the inviolability of raw data, define ownership of data and any confidentiality requirements and potential legal issues that need to be addressed as well. It would probably also be a good idea for the local partners to have similar documents drawn up.

Data Input:

There are inherent risks to the current system of partners inputting data into template Excel files, sending them over, the data being copied into other Excel files and then processed further. These have to do with conflating the data input, storage and processing aspects of the pipeline, with unnecessary extra steps, which increase the chance of error and with a lack of input controls.

MZ suggested looking into Google Forms as a preferred solution for data input. This would solve the following issues:

- we would prepare the form of raw data input, together with error checking
- partners input the data directly via the form
- input is separated from storage and processing
- only the admin (CP) has access to the stored raw data
- there is no manual copying of the data

In the event that the local partners have difficulty using Google Forms in the field they would still be a preferred solution, if only that meant that the data would have to be transferred from their spreadsheets manually by CP.

Data storage:

Raw data should be stored using a 'read-only' solution: so that data input is strictly controlled but that otherwise the data can only accessed in order to extract required subsets, but not to manipulate it in any way. The inviolability of the raw data can thereby be maintained in a manner that the current Excel solution does not guarantee: the data is safe from accidentally being overwritten or corrupted in any way. Using some form of version control also ensures that there is only one principal version of the data instead of several copies in different locations competing for primacy.

MZ suggests using Google Sheets as a preferred solution for data storage. This would solve the following issues:

- single authoritative version of data
- some degree of backup and version control (i.e. going back in time in case of data corruption)
- data can be protected and access clearly delimited

Data processing and data analysis:

Processing refers to simple derived variables such as rates or totals, or pivot tables, and analysis to more advanced methods including visualisation of data. We need to establish an exhaustive set of options for data processing and analysis that Chase Africa expects to use the data for.

MZ suggests using shiny as the platform to deploy a solution in the style of an interactive dashboard that would use the Google Sheet data directly as the raw data source. See here https://shiny.rstudio.com/gallery/ for examples of the possibilities shiny affords. But in brief shiny would allow:

- interactive visualisations of the data, including subsets and user defined variables of interest
- directly using the data from the Google Sheets file, without any danger of corruption
- easy extraction of subsets of the data table wihtout needing to use the raw data directly.
- the shiny app can be hosted on https://www.shinyapps.io/ for free up to 25 processing hours a month, which should suffice for a small user base such as in anticipated here (i.e. for internal use only). Even if that were ever surpassed the pricing is very reasonable (starter package is 100$ per year).

Although most of the discussion focused on the data management of the currently available (aggregate) data, we also discussed the possibility of gaining access to partner individual patient level data. There are many reasons to expect this will be a challenging endeavour, including the fact that these records are currently only kept in paper form, and the fact that the various partners record-keeping is not harmonized either. In these circumstances it will be difficult to change their practices and adding extra work to their load will likely be met with resistance, and any venture into collecting individual data should have a clear idea about the expected outcomes and the cost of such a project. Still, it is within the purview of this project to consider how all of the solutions outlined above could facilitate the future expansion of the data collection and analysis strategy. This applies in particular to the formalisation of the Data Management Plan, which should be written to be comprehensive and flexible.

Next steps:

1. Consolidating the existing data: The existing data needs to be consolidated using tidy data principles. To this end MZ will prepare the template tables for the merger of the existing data and CP will enter the existing data. We will manually perform checks to ensure this transition is error free.

2. Start drafting the Data Management Plan: CP will start drafting the DMP, with input from MZ on specific processes such as those discussed here.

3. Next meeting will probably take place in Oxford and will have the objective of setting up the Google Forms and Google Sheets by CP with MZ overseeing to ensure the capacity to establish this system and to adapt it if necessary in the future is retained by Chase Africa.

4. The next meeting will also have the objective to formalise the structure of the shiny app by considering the requirements of Chase Africa and the possibilities available on this platform.