## Triage purpose

This is an exercise which you will do alongside the data linkage enhancement team at the start of an engagement with the quality assurance framework for data linkage.

There are two objectives from this work:

* Firstly, it will start a conversation to identify potential linkage issues
* Secondly, this log forms an important part of the documentation alongside the framework

The first section of questions are multiple choice to establish some of the project’s core tenets, the second section asks more extended questions to ascertain a more in-depth view of the project’s risk profile.

## Part 1: Multiple choice triage questions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dimension** | **Question** | **1/0** | **Options** | **Example** | **Why is this question important?** | **Additional notes** |
| **Engagement start** | **What stage of your linkage project have you started engagement with the QAFDL?** |  | Start of the project | 111 linkage project | This affects the order with which elements of the QAFDL should be considered |  |
|  | Some project work has been completed, but work is ongoing | Echild |
|  | Linkage has been completed, engaging with QAFDL for audit purposes | PAVE |
| **Linkage project type** | **What project type does this engagement fall under?** |  | Type 1 | Creating a new data asset | Categorising projects in this way can help with transfer of lessons learnt, and signposting the most relevant documentation |  |
|  | Type 2 | Linking pre-existing assets to create a new asset (eg CVD prevent) |
|  | Type 3 | Auditing work which has already taken place (PAVE) |
|  | Type 4 | Assisting NHSE customers who have access to pseudonymised data and have no information on data linkag error rates (Our future health) |
| **Types of linkage algorithm** | Which type of linkage algorithm are you using? If you are not sure, carry out the Preparation phase on the QAF log to discover the best option for your project. Then, come back to this question and select the most appropriate answer. Are you constrained in using a pre-determined linkage algorithm (e.g. MPS)? | 1 | Simple join | Deterministic perfect match on NHS number and DOB | The data linker might need to use a specific, previously established methods for linking data, which affects how we check the quality of the linkage. |  |
|  | Established linkage algorithms like MPS | MPS |
|  | Established linkage algorithms + additional rules OR Ad-hoc models | Shielded Patients List, ECHILD, Vaccination Probabilistic model using Fellegi Sunter algorithm |
|  | Commissioned to third-parties | Open Safely (EMIS and TPP), ONS link PID to Census under DEA, HMRC, etc. |
| **Purpose** | For what purpose are the data sets being linked? |  | Direct care | Vaccination | The purpose influences the quality assurance needed, for example, linked data used for direct patient care requires the highest quality. Depending on whether the linkage is defining a cohort or defining attributes you will want to have differing levels of sensitivity and recall. |  |
|  | Non-direct care | NCMP (National Child Measurement Programme - links pupils data to PDS), ECDS and HES linkage (compare clinical events and complete the other assets where missing events are identified). |
| **Risk** | What is the risk associated with data linkage inaccuracies? (missing a link making the wrong one) |  | High risk | Those where incorrect data links can result in severe consequences in healthcare, potentially leading to harm, ethical or legal problems. | Understanding the risks involved allows us to adjust our data linkage approach to ensure accuracy and reliability for its intended use. |  |
|  | Moderate risk | Those where incorrect data links may have non-critical consequences. |
|  | Low-risk | Those where link errors will not significantly impact on patient care or research validity. |
| **Profile of data content** | What is the importance profile of the data ? |  | High profile health data | Support a strand of the long-term plan. High media attention. | The project's significance is determined by how errors could impact patient safety and may be influenced by other factors like business or politics. |  |
|  | Normal profile | Anything else |
| **Sensitivity of data** | How sensitive is the data we are using? |  | Sensitive data | Any disclosive information (patient clinical or demographics information) | We adjust our quality checks based on how sensitive the data is to protect it properly during linkage. |  |
|  | Non-sensitive data | Aggregated data that are non-disclosive of patients information Data publicly available on the internet |
| **Frequency** | How often do you plan on linking these data sets? |  | One-off | DARS cohort requests | The linkage frequency impacts how we approach the efficiency and repeatability of the process. |  |
|  | Repeated (yearly, monthly) | NCMP (National Child Measurement Programme) |
|  | Repeated (weekly, daily) | HES-ONS and all core assets |
|  | Live | Find my NHS Number |
| **Expertise** | How experienced is the person linking the data? |  | Intermediate+ | Have carried out 4 or more data linkage projects of different nature | The skill level of the person linking the data affects quality; our framework guides less experienced staff through the necessary steps. |  |
|  | Novice | Have carried out 3 or less data linkage projects (including the completion of the Quality Assurance Framework) |

## Part 2: Extended triage questions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Additional questions** | **Why is this question important?** | **Related QAFDL aspect** | **Notes** |
| **Entity being linked** | What entity is being linked (person, event, organisation,etc)? | Knowing what we're linking—people, events, or organisations—guides our choice of linking variables and the reference master dataset used. | Configuration of linkage parameters/settings |  |
| **Common characteristics used** | What common linkage fields do you have available in the two assets for linkage? (NHS numbers, MPS\_ID, pseudo (NHS number or MPS\_ID), DOB, etc | To link data successfully, we need to identify fields that are sufficient to identify a match. | Profiling |  |
| **Metadata** | Is metadata and documentation available for all datasets? | Having common linking fields might not be enough for a successful linkage. Understanding them is important to define how they need to be adjusted or derived. | Profiling |  |
| **Benchmarking** | What level of overlapping are you expecting from the data sets being linked? | Linked datasets may represent different groups, so knowing how much overlap to expect helps us to predict acceptable linkage rates. | Verification and validity |  |
| **Involvement of reference master data (a "spine")** | Do you need to link data sets to a reference master data set (a "spine") or between them or both? | Linking to a main reference dataset, or "spine," is best practice. It helps to connect different datasets and track changes over time, like when a patient moves house. | Techniques and tools |  |
| **Synchronicity of data collection** | Do the data sets being linked refer to the same time frame? And are the PIDs being collected at the same time than the analytical measures of interest? | If datasets span different times, we must adjust for these differences and consider how they might affect the linking process, potentially evaluating how errors in linkage change over time. | Profiling, assessment |  |
| **Source of data** | What categories do the data sets being linked fall into? How was the data collected? Is it clinically validated, or patient inputted? (e.g. Cohort from external researchers, Core assets, DARS assets, Scraped data , Collected by NHSE, Sent by OGD) | This context can provide lots of clues as to how best to link, what kind of quality we are expecting with respect to outputs, and what kind of quality assurance is needed post linkage. The source of data used in the linkage process can support decisions around other dimensions. For example, core assets are usually linked to reference master data sets via pipelines set up and maintained by the data engineering team. | Data access |  |
| **Data types** | What data types are involved in the linkage (numerical, free text, binary, strings, etc)? | The nature of the data we link (numbers, text, yes/no answers) affects how we prepare it for linking and which tools we choose. Data types need to be consistent across linkage variables between datasets. | Profiling |  |
| **Data quality** | What is the expected data quality of the data sets being linked? (high, low) | Data quality drives our choice of linking methods—better quality data allows for simpler linking techniques. | Assessment |  |
| **Number of data sets** | How many data sets will be linked? | When linking several datasets, we must consider how they'll group together and how to resolve any mismatches, shaping our evaluation process. | Configuration of linkage parameters/settings |  |
| **Environments where the linkage is executed** | Are you constrained in which environment will you be executing the linkage (e.g. DAE, DPS, UDAL, NDSR, RDS, FDP, etc)? | The setting where we link data can dictate what data fields, tools, and methods we can use, possibly restricting our options of linkage algorithm. | Computational resources, Safety, Configuration of linkage parameters/settings, Techniques and tools |  |
| **Environments where the assets reside** | Where do the assets reside (DAE, UDAL, RDS, etc)? | Knowing the location of our data helps anticipate and plan for any hurdles, like secure transfers and privacy protection needs. | Safety |  |
| **Customer request** | Is the customer asking for the highest level of quality assurance regardless of the triage outcomes? | A project might be commissioned by a customer that requires the highest level of quality assurance regardless of the project characteristics.  The opposite may also occur, however, it is always recommended that the minimum suggested level of quality assurance is delivered to guarantee a professional outcome. | Knowledge management, Quality of linkage |  |
| **Resources** | Which resources (and what type) do you have available to carry out the linkage? | Our desired quality level might be limited by available resources, and we need to be clear when we can't meet our ideal standards. | Uncertainty management |  |