MPS Diagnostics User Guide

Interpretable metadata for MPS and Person ID

Document Name	MPS Diagnostics User Guide		
Project / Programme	**MPS Enhancement**	Project	**MPS Enhancement**
Project Manager	**Giulia Mantovani**	Status	**Final**
Owner	**Rupert Chaplin**	Version	**1.0.0**
Author	**Data Science Team**	Version issue date	**16/11/2023**

Document management

Revision History

Version	Date	Summary of Changes
0.0.1	22/06/2023	First draft
1.0.0	16/11/2023	Published

Reviewers

This document must be reviewed by the following people:

Reviewer name	Title / Responsibility	Date	Version
Rupert Chaplin	Head of Data Science	13/11/2023	1.0.0

Approved by

This document must be approved by the following people:

Name	Signature	Title	Date	Version
Rupert Chaplin		Head of Data Science	13/11/2023	1.0.0

Document Control:

The controlled copy of this document is maintained in the NHS England corporate network. Any copies of this document held outside of that area, in whatever format (for example paper, email attachment), are considered to have passed out of control and should be checked for currency and validity.

Contents

Executive summary

The *Person_ID* is a unique patient identifier used by NHS Digital with the objective of standardizing the approach to patient-level data linkage across different data sets.

*Person_ID*s are provided in many data sets available in NHS Digital, and are derived by the Master Person Service (MPS). For security and privacy reasons many users might have visibility of the tokenised version of the *Person_ID*, which provides an extra level of patient confidentiality.

MPS takes certain patients' demographic information, and matches it to their record (associated with a unique *NHS number*) in the Personal Demographics Service (PDS) data set. PDS is the collection of all *NHS number*s and patients' demographic information.

An accompanying document, [The Person_ID handbook for HES users](#), explains in detail the algorithms by which MPS derives a *Person_ID* from the demographic information of the patient. In that document, we gave several example case studies where the *Person_ID* returned by MPS may confuse analysts. Such confusion can occur because analysts are presented with only the *Person_ID*, without contextual information about how the *Person_ID* was derived, for example which demographic data led to finding the match. Some relevant contextual information can be found in the MPS response record (returned by MPS), however this record is not usually shared with users.

MPS Diagnostics is a new pipeline which takes the contextual information from the MPS response file, and some additional data from PDS, to create 10 columns of metadata explaining in user-friendly terms how each *Person_ID* was derived. MPS Diagnostics pipeline generates the mps_diagnostics data set, which contains record identifiers and the MPS diagnostics columns.

mps_diagnostics is available upon request for internal NHS England analysts via CDAs (clear data agreements), or for external NHS E users via DSAs (data sharing agreements). This document details the 10 columns, how they are derived and how they are to be used.

Abbreviations and glossary

The following table contains a quick reference for abbreviation terms across the document.

Term	Definition	Meaning
DARS	Data Access Request Service	The Data Access Request Service can offer clinicians, researchers and commissioners the data required to help improve NHS services. See DARS.
DPS	Data Processing Services	The technologies and processes which enable NHS England to collect, process, and access data. See DPS.
MPS	Master Person Service	A system that enables data linkage between data sets through Person_ID by tracing and verifying key identifiers against the Personal Demographics Service (PDS). See MPS.
NBO	National Back Office	The NBO provides a national data quality service. It is responsible for the management of NHS numbers and PDS records, investigation and resolution of data quality incidents on PDS demographic records, and the provision of a Tracing Service to approved organisations. See NBO.
PDS	Personal Demographic Service	PDS is the national electronic database of NHS patient details, which holds about 80M records. The demographic details are normally updated when patients visit their GP but may also be updated by other healthcare professionals and from other sources See PDS.
UPRI1	Unmatched person record identifier 1	Another term used for MPS_ID – the identifier generated if a match cannot be found in PDS but a record can be linked or created in the MPS record bucket.
UPRI2	Unmatched person record identifier 2	This is the identifier generated if neither an NHS number nor MPS_ID (UPRI 1) exists.

The following table contains a quick reference for common terms across the document.

Term	Meaning
NHS number	The *NHS number* is a unique identifier for a patient within the NHS in England and Wales. The specific data fields for *NHS number*s might have different names in different data sets (for example, in HES it is called *NEW_NHS_NO*). See <u>NHS number</u>.
Person_ID	The *Person_ID* can be either the *NHS number,* if a match for the record was found in PDS, or the *MPS_ID* if the record was matched in the MPS record bucket or sufficient demographic information was provided to create a new *MPS_ID*. If neither an *NHS number* nor an *MPS_ID* were provided, the record is assigned a *one-time-use* unique identifier as *Person_ID*.
MPS_ID	Identifier returned by MPS containing the unique identifier to records in the MPS record bucket.
MPS record bucket	Records which cannot be matched in PDS but contain sufficient demographic information for a match are stored in the MPS record bucket. This is so that future instances of the same or similar record can be assigned a consistent ID, the *MPS_ID,* despite not matching to PDS.
Algorithmic trace	Algorithmic trace is the final stage in MPS to match records with PDS. For each query record, a set of PDS candidate records are identified by blocking. Each PDS candidate is then scored. The highest scoring candidate record is chosen as the matching record.
Alphanumeric trace	Alphanumeric trace is the second trace step in MPS. The minimum required fields are family name, year of birth and gender, but as HES does not contain names, this step is skipped for HES data sets.
Cross-check trace	Cross-check trace is the first and simplest tracing step, which can be used when NHS number and DOB are present. This appears twice, once against the cached version of PDS within DPS and a more complex version is present in Spine against the live version of PDS.
Spine	Spine supports the IT infrastructure for health and social care in England, joining together healthcare IT systems. See <u>Spine</u>.

1. Introduction

1.1   MPS overview

MPS takes certain demographic information, such as that contained in a person's health and care records, and matches it to their unique *NHS number* to confirm their identity. The collection of all *NHS number*s and patients' demographic information is contained in the Personal Demographics Service (PDS) data set.

Where a perfect match of *NHS number* and *date of birth* cannot be found between a record of interest and any of the PDS records, more complex algorithms are used to compare partial demographic information and identify the most likely PDS record corresponding to the query record. These algorithms are referred to as alphanumeric and algorithmic trace steps. The alphanumeric trace step looks for a single exact match in PDS based on three mandatory demographics, that is, *family name*, *year of birth*, and *gender*, and some other demographic variables such as postcode, GP practice and others if available. In the algorithmic trace step, the single queried record is compared to all records in PDS. The comparisons involve demographic information (*date of birth*, *family name*, *given name*, *gender* and *postcode*) and are scored based on similarity. If the similarity is deemed acceptable and there is no ambiguity between multiple PDS records (that is, there is a match in PDS that is clearly better than other possible matches), the best matched record is returned. Otherwise, the algorithm proceeds to look for similarities between the record of interest and some previously unmatched records, stored in the MPS record bucket, a separate data set.

MPS_ID is also sometimes known as UPRI 1 (unmatched person record identifier 1). If the queried record could return neither an NHS number, nor an MPS_ID, then a *one-time-use ID* is generated, sometimes known as UPRI 2 (unmatched person record identifier 2). The *Person_ID* is therefore one of *NHS number* from PDS, *MPS_ID* from the MPS record bucket or a *one-time-use ID*, depending on if and where a match was found.

1.2   Why do we need MPS Diagnostics?

A goal of data linkage is for transparency in results. In requirements gathering sessions with users of data linked via MPS, some scenarios were explored where the *Person_ID* matched by MPS caused confusion and sometimes reduced the level of trust in the linkage approach. Two such examples are explained here. With the additional interpretable metadata provided by the MPS Diagnostics pipeline, users will have a way of identifying these scenarios, getting an insight on how the matching process happened, and thus regaining trust in the matching process.

Superseded NHS numbers

An *NHS number* can be superseded in PDS, which means that it is no longer valid, and it has been replaced by another one. If a query record contains a superseded *NHS number*, MPS can recognize this and return the corresponding valid *NHS number*. This can be confusing for analysts. For example, they may know that the cross-check trace step of MPS requires an exact match on *NHS number*, however because of superseded *NHS number*s, MPS might return a *Person_ID* that is different from the submitted NHS number, even at cross-check trace.

One-time-use IDs and longitudinal patient records

The *Person_ID* returned by MPS can be an *NHS number*, an *MPS_ID*, or a *one-time-use ID*. In the case of *NHS number*, this means that a match is found for the query record in PDS. *MPS_ID*s instead are not found in PDS but generated from the MPS record bucket. And lastly, *one-time-use ID*s are instead generated when the query record could not be matched in either PDS, nor MPS record bucket, and does not have sufficient demographic information in it to generate a new MPS record bucket record either, essentially meaning that these records are not traceable in the future.

When analysts have access to clear data, they can identify what type of *Person_ID* they have received. *NHS numbers* are 10 digits, *MPS_ID*s begin with "A" or "B", and *one-time-use ID*s begin with "U". However, users with access to pseudonymised data have no way to distinguish between these types of *Person_ID*.

This causes a particular problem when trying to create a longitudinal patient record. This will not be possible for patients for whom *one-time-use ID*s were generated, so analysts may choose to exclude *one-time-use ID*s from the scope of such studies. Similarly, analysis on re-admission rate, or rate of repeated use of a service, can be biased by patients with *one-time-use ID*s, especially for some parts of the population systematically affected by lower data quality, so analysts may choose to exclude *one-time-use ID*s.

1.3   Scope of this document

Chapter 2 explains the 10 metadata columns of MPS Diagnostics, how they are derived and how they can help analysts to better understand the *Person_ID* returned by MPS.

Chapter 3 describes the schema of MPS Diagnostics, including tokenised (pseudonymised) fields.

Chapter 4 explains how MPS Diagnostics can be joined with original datasets.

Chapter 5 gives an example aggregated report that analysts would find a useful summary of MPS Diagnostics metadata, to explain how MPS has matched their records.

Chapter 6 revisits the case studies from The Person_ID handbook for HES users, showing where the metadata from MPS Diagnostics would have helped to make the interpretation of the results more clear.

2. MPS Diagnostics definitions and derivations

2.1 Person ID type

Definition

The type of *Person_ID* returned by MPS. This could be:

- NHSNUMBER: if a matching record was found in PDS.

- MPS_ID: if a matching record was found in the MPS record bucket, or sufficient information was present in the query record to generate a new *MPS_ID*. This is also known as UPRI1.

- ONE_TIME_USE_ID: if the record could not be matched to PDS, nor to the MPS record bucket, and sufficient information was not present in the query record to generate a new *MPS_ID*. Then a *one-time-use ID* is generated. This is also known as UPRI2.

Derivation

- NHSNUMBER is returned if the *Person_ID* consists of 10 digits

- MPS_ID is returned if the *Person_ID* begins with "A" or "B"

- ONE_TIME_USE_ID is returned if the *Person_ID* begins with "U".

These derivation rules are applied to the clear (not pseudonymised) *Person_ID*. The value of Person ID type is that even if the user is in receipt of pseudonymised data, they will still be able to tell from the Person ID type whether a match was found in PDS or the MPS record bucket, or neither.

2.2 MPS last step attempted

Definition

The last trace step attempted by MPS. This could be:

- CCT_cached: if the match was found by cross-check trace in DPS

- CCT_live: if the match was found by cross-check trace in Spine

- alphanumeric_trace_live: if the match was found by alphanumeric trace in Spine

- algorithmic_trace_live: if the match was found by algorithmic trace in Spine

- No_PDS_tracing_run: if none of the tracing steps were run.

Derivation

Derived from fields in the MPS response file:

- CCT_cached: if *MatchedAlgorithmIndicator*=1 and *DateOfBirthScorePercentage*=null

- CCT_live: if *MatchedAlgorithmIndicator*=1 and *DateOfBirthScorePercentage*=0

- *alphanumeric_trace_live*: if *MatchedAlgorithmIndicator*=3

- *algorithmic_trace_live*: if *MatchedAlgorithmIndicator*=4

- *No_PDS_tracing_run*: if *MatchedAlgorithmIndicator*=0

MatchedAlgorithmIndicator and *DateOfBirthScorePercentage* are fields returned in the MPS response record. Full explanation of these, and of the MPS trace steps can be found in The Person_ID handbook for HES users.

MPS last step attempted gives users insight additional to what they learn from MPS successful step (next), in the case where no MPS trace steps were successful, users will be able to see if, for example, algorithmic trace was skipped (which would be due to the query record not containing the minimum fields required for algorithmic trace).

2.3   MPS successful step

Definition

The trace step at which MPS has successfully matched to a record in PDS. This could be:

- CCT_cached: if the match was found by cross-check trace in DPS. This means that the record was matched to PDS with an exact match on *NHS number* and *date of birth*.

- CCT_live: if the match was found by cross-check trace in Spine. This means that the record was matched to PDS with an exact match on *NHS number*, a partial match on *date of birth*, and either a match on names, or a match on the first part of the *postcode*. Alternatively, it matched perfectly on both *NHS number* and *date of birth* but the *NHS number* had been superseded in PDS. This use case is not captured in CCT_cached.

- alphanumeric_trace_live: if the match was found by alphanumeric trace in Spine. This means that records were predominantly matched with *family name*, *year of birth* and *gender*. Other personal information may have also been used during the tracing. No *NHS number* was used at this stage, and the match needs to be perfect on the demographic information used by the trace.

- algorithmic_trace_live: if the match was found by algorithmic trace in Spine. This means that there was an exact or partial match on at least three of these five fields: *family name*, *given name*, *gender*, *date of birth*, *postcode*. See The Person_ID handbook for HES users for precise explanation of these and the other matching rules. Users can also refer to the "MPS algorithmic match score" (chapter 0) to find out which of these five fields were used to match.

- No_PDS_tracing_run: if none of the tracing steps were run.

- No_PDS_match_found: if some or all trace steps were run but did not find a match in PDS.

Derivation

Derived from fields in the MPS response file:

- CCT_cached: if *MatchedAlgorithmIndicator*=1, *DateOfBirthScorePercentage*=null, and *MatchedConfidencePercentage*=100

- CCT_live: if *MatchedAlgorithmIndicator*=1, *DateOfBirthScorePercentage*=0, and *MatchedConfidencePercentage*=100

- alphanumeric_trace_live: if *MatchedAlgorithmIndicator*=3 and *MatchedConfidencePercentage*=100

- algorithmic_trace_live: if *MatchedAlgorithmIndicator*=4 and *MatchedConfidencePercentage >= 5*0

- No_PDS_tracing_run: if *MatchedAlgorithmIndicator*=0

- No_PDS_match_found: if *MatchedAlgorithmIndicator* != 0, and *MatchedConfidencePercentage*=0

2.4   PDS match flag

Definition

This indicates whether a match was found in PDS. In can be either True or False.

- True: the match was found in PDS via CCT_cached, CCT_live, alphanumeric_trace_live, or algorithmic_trace_live

- False: no match was found on PDS.

Derivation

Derived from Person ID type (chapter 2.1).

- True: if Person ID type = NHSNUMBER

- False: otherwise.

This should not be interpreted as a guarantee that the match found on PDS is correct, only that a match was found within the expected levels of errors of MPS.

2.5   Superseded NHS number flag

Definition

This indicates whether the NHS number returned by MPS superseded the NHS number submitted in the query record.

- True: the NHS number returned by MPS superseded the NHS number submitted in the query record

- False: otherwise (note, the *NHS number* submitted in the query record might still be superseded on PDS, but not by the specific *NHS number* returned by MPS).

Derivation

Derived from PDS:

- True: if the PDS record returned by MPS contained in its *replacementOf* field the *NHS number* from the query record, at the time of the query record's *AS_AT_date*

- False: otherwise.

This derivation makes use of the "*from*" and "*to*", and "*sedLow*" and "*sedHigh*" dates in PDS. ("sed" means system effective date) to ensure that the Superseded NHS number flag is derived consistently, regardless of when it is derived, since any temporal dependency is based on the query's *AS_AT_date*.

As described in chapter 1.2, when MPS returns an *NHS number* which is different to the one in the query record, it can be confusing for users, especially when the trace step was CCT_live, which requires an exact match on *NHS number*. Having the superseded flag will clear up this confusion.

2.6   NHS number history list

Definition

List of other *NHS number*s associated with the same patient.

PDS is curated by the National Back Office team (NBO). When NBO find more than one record which appears to be the same patient, they choose one which replaces/supersedes the others. This is managed via the *replacedBy* and *replacementOf* columns in PDS. This process does not include when a record has been invalidated due to legal or safeguarding reasons (for example, adoption, gender reassignment). In such cases, the *replacedBy* and *replacementOf* columns are not used.

For example:

NHS number	replacedBy	replacementOf
1111111111	2222222222	null
2222222222	3333333333	1111111111
3333333333	null	1111111111, 2222222222

FICTICIOUS DATA

Here, NBO discovered that records 1111111111 and 2222222222 appear to be the same patient, so they marked 1111111111 as being replaced by 2222222222. Later, they discovered that record 3333333333 also appears to be the same patient, so they marked 2222222222 as being replaced by 3333333333.

Hence, the full NHS number history list for 3333333333 would be [1111111111, 2222222222].

Derivation

Derived from PDS:

- The list of *NHS number*s found in the *replacementOf* field of the PDS record returned by MPS, at the time of the query record's *AS_AT_date*. The list is in order of the *replacementOf from* date in PDS.

The NHS number history list enables analysts to create a complete longitudinal patient record, even when a patient has been associated with multiple *NHS number*s.

How to use NHS number history list when *study_IDs* are provided.

Oftentimes NHS Digital is provided with cohorts from research studies that need to be linked to other data assets (for example, HES). For data minimisation reasons, such cohorts might only contain a *study_ID* to identify the patient. When the cohort is linked to HES via the *Person_ID* (or *NHS number*) obtained from MPS, it is responsibility of the person that carries out the linkage to utilise NHS number history list correctly and map the *study_ID* to the multiple *NHS numbers* returned by NHS number history list to make sure that the complete longitudinal patient record is returned.

2.7   Multiple PDS matches flag

Definition

This flag indicates that MPS did not return an *NHS number* due to there being multiple plausible matches. For example, at the algorithmic trace step there could be more than one match in PDS to the demographic information in the query record, and the scoring on those matches is equal or close to equal.

- True: MPS found multiple plausible matches on PDS, therefore returned none of them

- False: otherwise.

Derivation

Derived from the MPS response record:

- True: if the MPS response record contained the error code "97", indicating multiple matches
- False: otherwise.

The Multiple PDS matches flag can help if rich demographic data has been sent in the query record, but no match is found on PDS. This flag explains that the lack of a match is not due to insufficient data being submitted.

2.8   Multiple MPS ID matches flag

Definition

If MPS has not returned an *NHS number*, it could return an *MPS_ID* from the MPS record bucket. Sometimes there are multiple equally plausible matches in the MPS record bucket, and only one is chosen to populate the *Person_ID*.

- True: there were multiple matches and the *Person_ID* returned is only one of many plausible *MPS_ID*s
- False: otherwise.

Derivation

Derived from the MPS response record:

- True: if the *MPS_ID* contains more than one *MPS_ID* (strings separated by "~~~")
- False: otherwise.

2.9   MPS match score

Definition

The score returned by MPS indicating the quality of the match. These scores are calculated differently depending on which trace step was successful. At cross-check trace and alphanumeric trace it is always 100. At algorithmic trace it is a score between 50 and 100, which is aggregated from the five MPS algorithmic match scores (see chapter 2.10).

- Possible values: Decimal between 0 and 100.

The MPS match score does not represent a probabilistic measure of confidence.

Derivation

Derived from the MPS response record:

- MPS match score is a copy of the *MatchedConfidencePercentage* in the MPS response record.

2.10 Algorithmic match score

Definition

When MPS finds a match using algorithmic trace, these columns contain the match score for each of the demographic fields: *family name*, *given name*, *date of birth*, *gender*, and *postcode*.

- Column names:
    - FAMILY_NAME_ALGORITHMIC_MATCH_SCORE
    - GIVEN_NAME_ALGORITHMIC_MATCH_SCORE
    - DATEOFBIRTH_NAME_ALGORITHMIC_MATCH_SCORE
    - GENDER_ALGORITHMIC_MATCH_SCORE
    - POSTCODE_ALGORITHMIC_MATCH_SCORE
- Possible values: Each column contains a Decimal between 0 and 100.

Derivation

Derived from the MPS response record:

- MPS algorithmic match score is a copy of these fields from the MPS response record:
    - *ALGORITHMICTRACE_FAMILY_NAME_SCORE_PERC*
    - *ALGORITHMICTRACE_GIVEN_NAME_SCORE_PERC*
    - *ALGORITHMICTRACE_DOB_SCORE_PERC*
    - *ALGORITHMICTRACE_GENDER_SCORE_PERC*
    - *ALGORITHMICTRACE_POSTCODE_SCORE_PERC*

This enables the analyst to see which of the demographic information in the query record was used successfully in finding the match on PDS.

3. Schema of mps_diagnostics

Column name	Type	Category	Description	Allowable values
dataset_id	string	Record identifier	The dataset the record belongs to	any
local_id	string	Record identifier	An identifier for the batch of records which were sent to MPS	any
unique_reference	string	Record identifier	An identifier for the record which was sent to MPS. This is unique with respect to the Local ID.	any
req_created	timestamp	MPS request and response record field	The date on which the record was sent to MPS.	datetime
req_AS_AT_DATE	date	MPS request and response record field	The date the data sent to MPS was valid for.	date
res_created	timestamp	MPS request and response record field	The date on which MPS generated the response to the request record.	datetime
res_AS_AT_DATE	date	MPS request and response record field	The date the response from MPS was valid for.	date
PERSON_ID	string	MPS request and response record field	The PERSON_ID found or generated by MPS	10 characters
MATCHED_NHS_NO	string	MPS request and response record field	The NHS number found by MPS	10 digits

PERSON_ID_TYPE	string	MPS Diagnostics metadata field	Whether the Person_ID is an NHS number, an MPS_ID (UPRI1), or a one-time use ID (UPRI2)	NHSNUMBER MPS_ID ONE_TIME_USE_ID
MPS_LAST_STEP_ATTEMPTED	string	MPS Diagnostics metadata field	The last trace step attempted by MPS	CCT_cached CCT_live alphanumeric_trace_live algorithmic_trace_live No_PDS_tracing_run
MPS_SUCCESSFUL_STEP	string	MPS Diagnostics metadata field	The MPS trace step which successfully matched an NHS number	CCT_cached CCT_live alphanumeric_trace_live algorithmic_trace_live No_PDS_tracing_run No_PDS_match_found
PDS_MATCH_FLAG	boolean	MPS Diagnostics metadata field	Flags whether an NHS number was successfully matched from PDS	True / False
SUPERSEDED_NHS_NUMBER_FLAG	boolean	MPS Diagnostics metadata field	This flag indicates whether the NHS number in the MPS request was superseded by the NHS number in the MPS response	True / False
NHS_NUMBER_HISTORY_LIST	array<string>	MPS Diagnostics metadata field	Lists other NHS numbers associated with the patient matched by MPS	List of 10 digit numbers
MULTIPLE_PDS_MATCHES_FLAG	boolean	MPS Diagnostics metadata field	This flag indicates that no NHS number was returned by MPS because there were multiple	True / False

			records in PDS matching the request, rather than due to insifficient or low quality data	
MULTIPLE_MPS_ID_ MATCHES_FLAG	boolean	MPS Diagnostics metadata field	This flag indicates that multiple MPS_IDs were returned by MPS	True / False
MPS_MATCH_SCOR E	decimal(5,2)	MPS Diagnostics metadata field	The MPS match score calculated by MPS. This calculation is different depending on which MPS trace step was successful	Decimal between 0 and 100
FAMILY_NAME_ALG ORITHMIC_MATCH_ SCORE	decimal(5,2)	MPS Diagnostics metadata field	When a match was found at the algorithmic trace step of MPS, this contains a match score for family name	Decimal between 0 and 100
GIVEN_NAME_ALGO RITHMIC_MATCH_S CORE	decimal(5,2)	MPS Diagnostics metadata field	When a match was found at the algorithmic trace step of MPS, this contains a match score for given name	Decimal between 0 and 100
DATEOFBIRTH_ALG ORITHMIC_MATCH_ SCORE	decimal(5,2)	MPS Diagnostics metadata field	When a match was found at the algorithmic trace step of MPS, this contains a match score for date of birth	Decimal between 0 and 100
GENDER_ALGORITH MIC_MATCH_SCORE	decimal(5,2)	MPS Diagnostics metadata field	When a match was found at the algorithmic trace step of MPS, this contains a match score for gender	Decimal between 0 and 100
POSTCODE_ALGORI THMIC_MATCH_SCO RE	decimal(5,2)	MPS Diagnostics metadata field	When a match was found at the algorithmic trace step of MPS, this contains a match score for postcode	Decimal between 0 and 100
MPS_DIAGNOSTICS _TIMESTAMP	timestamp	MPS Diagnostics metadata field	The date on which the MPS diagnostics columns were derived.	datetime

Tokenised columns

In some data sharing agreements, columns containing *NHS number*s and *Person_ID*s must be pseudonymised via tokenisation. In such agreements, some of the following columns will be available instead of their clear version.

Column name	Type	Description	Allowable values
TOKEN_ PERSON_ID	string	Pseudonymised identifier for Person_ID.	any
TOKEN_MATCHED_NHS_NUMBER	string	Pseudonymised identifier for the NHS number found by MPS.	any
TOKEN_NHS_NUMBER_HISTORY_LIST	string	Pseudonymised identifiers for the NHS number history list. Pseudonymised such that they could be joined to other pseudonymised NHS numbers.	any
TOKEN_NHS_NUMBER_HISTORY_LIST_AS_PERSON_ID	string	Pseudonymised identifiers for the NHS number history list. Pseudonymised such that they could be joined to other pseudonymised Person_IDs.	any

Watch out for the difference between tokenised *NHS numbers* and *Person_IDs.*

A tokenised *NHS number* is different to a tokenised *Person_ID*, even if the clear *NHS number* and *Person_ID* hold the same value. Therefore, analysts should decide whether to use *NHS number* columns tokenised as *NHS number* or tokenised as *Person_ID*, depending on whether they intend to join it to data with *NHS number*s or *Person_ID*s.

The table below helps explaining such situation. Joining *NHS numbers* to *Person_IDs* when having access to data in the clear will still work because the *Person_ID* is equal to the *NHS number* (when the *NHS number* was found essentially). However, when joining on tokenised values, this would not work because the two fields are tokenised differently.

Notably, this does not constitute a problem for *MPS_IDs* or *one-time-use IDs*.

NHS_NUMBER	PERSON_ID	TOKEN_NHS_NUMBER	TOKEN_PERSON_ID
0123456789	0123456789	227G114447654A111198II2223333888	987A543219876G432198RR5432198765
-	A123456789	-	647A5142198RRT432198RHH432112332
-	U123456789	-	1017JJ146HH8R12322198RHYY771KK32

4. How to join mps_diagnostics records to their original assets

Record identifier columns

Records in mps_diagnostics have three record identifier columns so that analysts can join the mps_diagnostics information back to their original asset.

- *dataset_id*: the name of the dataset. For example "AHAS", for the HES AHAS dataset

- *local_id*: this refers to the batch of records that was sent to MPS

- *unique_reference*: the unique reference with respect to the local_id.

Records in mps_diagnostics are unique on local_id and unique_reference.

Joining to original assets

Users of assets which have been MPS-checked (that is, linked to PDS via MPS) as part of their pipeline will be able to join their records to mps_diagnostics by joining their primary key to the "unique_reference". For example, users of HES AHAS will find that *unique_reference* corresponds to their "EPIKEY" (episode key), "AEKEY" (A&E key), or "ATTENDKEY" (attendance key). It is not necessary for users to know the *local_id*, as this is not visible on the HES AHAS record.

In very rare cases, the HES AHAS pipeline may have sent the same batch of records to MPS more than once. In such cases, mps_diagnostics has been disambiguated on *local_id* and *unique_reference*, keeping the most recent record (according to MPS response date) when there are duplicates.

Where assets use MPS as a service (MPSaaS), it is not guaranteed that *unique_identifier*s are indeed unique, because they are set by users. However, the combination of *local_id* and *unique_reference* is unique. The *local_id* is not known to users, so they may need to identify the batch of records based on the date they sent it and the size of the batch.

In rare cases, there may be multiple records in a batch with the same *local_id* and *unique_identifier*, This indicates that a user submitting a batch to MPSaaS has used the same primary key for more than one record. These ambiguous records have all been removed from mps_diagnostics, to avoid the risk of an analyst joining back to the wrong record.

Person_ID column

mps_diagnostics has a *Person_ID* column, which can be used as additional verification that a record in mps_diagnostics corresponds to a record in the original asset.

However, analysts should note that in the case of *one-time-use IDs*, the *Person_ID* column in mps_diagnostics is null. This is because *one-time-use ID*s are generated by DPS, after receiving MPS response records. mps_diagnostics is derived mostly from MPS Request and Response files, so does not know the *one-time-use ID*s. Analysts will be able to tell when this is the case by seeing a null entry in the *Person_ID* column and by seeing "ONE_TIME_USE_ID" in the *PERSON_ID_TYPE* column.

5. Aggregated reports

Some analysts may prefer to receive aggregated reports rather than record level MPS Diagnostics metadata. This can help them understand at a summary level why their records matched or did not match with MPS.

For example:

PERSON ID TYPE	MPS SUCCESSFUL STEP	MPS LAST STEP ATTEMPTED	PDS MATCH FLAG	MULTIPLE PDS MATCHES FLAG	MULTIPLE MPS ID MATCHES FLAG	SUPERSEDED NHS NUMBER FLAG	count	explanatory notes for customers
NHSNUMBER	CCT cached	CCT cached	TRUE	FALSE	FALSE	FALSE	695206	These would be the records matched on DPS. They were matched to PDS on NHS number and exact date of birth.
NHSNUMBER	CCT live	CCT live	TRUE	FALSE	FALSE	FALSE	1586	Found in PDS by cross check trace live. Exact match on NHS number, partial match on date of birth, and exact match on names or outcode (first part of postcode)
NHSNUMBER	CCT live	CCT live	TRUE	FALSE	FALSE	TRUE	3381	Found in PDS by cross check trace live. The NHS number returned is different to the NHS number submitted, because it was superseded. The full list of NHS numbers associated with this patient can be found by exploring the "NHS number history list" field (available as record-level info, not shown in this summary table). Exact match on NHS number, partial or exact match on date of birth, and exact match on names or outcode (first part of postcode)
NHSNUMBER	alphanumeric trace live	alphanumeric trace live	TRUE	FALSE	FALSE	FALSE	333	Found in PDS by alphanumeric trace. Matched on mandatory fields family name, year of birth, and gender (other demographic info could have been additionally used). NHS number is NOT used in this step.

PERSON ID TYPE	MPS SUCCESSFUL STEP	MPS LAST STEP ATTEMPTED	PDS MATCH FLAG	MULTIPLE PDS MATCHES FLAG	MULTIPLE MPS ID MATCHES FLAG	SUPERSEDED NHS NUMBER FLAG	count	explanatory notes for customers
NHSNUMBER	algorithmic trace live	algorithmic trace live	TRUE	FALSE	FALSE	FALSE	493	Found in PDS by algorithmic trace. Matched on at least 3 of these 5 demographic fields: family name, given name, date of birth, gender, postcode. Detail of which demographic fields matched can be found by exploring the "Algorithmic Match Score" fields (available as record-level info, not shown in this summary table)
NHSNUMBER	algorithmic trace live	algorithmic trace live	TRUE	FALSE	FALSE	TRUE	35	Found in PDS by algorithmic trace. The NHS number returned is different to the NHS number submitted, because it was superseded. The full list of NHS numbers associated with this patient can be found by exploring the "NHS number history list" field (available as record-level info, not shown in this summary table). Matched on at least 3 of these 5 demographic fields: family name, given name, date of birth, gender, postcode. Detail of which demographic fields matched can be found by exploring the "Algorithmic Match Score" fields (available as record-level info, not shown in this summary table)
MPS_ID	No PDS match found	CCT live	FALSE	FALSE	FALSE	FALSE	468	Not found in PDS, but matched against an identifier from the MPS record bucket. The last trace step attempted was cross check trace, because there was not sufficient valid demographic information submitted to attempt alphanumeric or algorithmic trace.

PERSON ID TYPE	MPS SUCCESSFUL STEP	MPS LAST STEP ATTEMPTED	PDS MATCH FLAG	MULTIPLE PDS MATCHES FLAG	MULTIPLE MPS ID MATCHES FLAG	SUPERSEDED NHS NUMBER FLAG	count	explanatory notes for customers
MPS_ID	No PDS match found	CCT live	FALSE	FALSE	TRUE	FALSE	30	Not found in PDS, so an ID from the MPS record bucket was returned. The last trace step attempted was cross check trace, because there was not sufficient demographic information submitted to attempt alphanumeric or algorithmic trace. There were multiple matches found in the MPS record bucket, only one of which was returned as the Person_ID.
MPS_ID	No PDS match found	alphanumeric trace live	FALSE	FALSE	FALSE	FALSE	31	Not found in PDS, but matched against an identifier from the MPS record bucket. The last trace step attempted was alphanumeric trace, because there was not sufficient valid demographic information submitted to attempt algorithmic trace.
MPS_ID	No PDS match found	alphanumeric trace live	FALSE	FALSE	TRUE	FALSE	0	Not found in PDS, but matched against an identifier from the MPS record bucket. The last trace step attempted was alphanumeric trace, because there was not sufficient valid demographic information submitted to attempt algorithmic trace. There were multiple matches found in the MPS record bucket, only one of which was returned as the Person_ID.
MPS_ID	No PDS match found	algorithmic trace live	FALSE	FALSE	FALSE	FALSE	6175	Not found in PDS, but matched against an identifier from the MPS record bucket. All the previous tracing steps were attempted because sufficient valid demographic information was submitted.
MPS_ID	No PDS match found	algorithmic trace live	FALSE	FALSE	TRUE	FALSE	96	Not found in PDS, but matched against an identifier from the MPS record bucket. There were multiple matches found in the MPS record bucket, only one of which was returned as the Person_ID.

PERSON ID TYPE	MPS SUCCESSFUL STEP	MPS LAST STEP ATTEMPTED	PDS MATCH FLAG	MULTIPLE PDS MATCHES FLAG	MULTIPLE MPS ID MATCHES FLAG	SUPERSEDED NHS NUMBER FLAG	count	explanatory notes for customers
ONE_TIME_USE_ID	No PDS match found	CCT live	FALSE	FALSE	FALSE	FALSE	1274	Not found in PDS, or in the MPS record bucket, so the Person_ID is a one-time-use ID. The last trace step attempted was cross check trace, because there was not sufficient valid demographic information submitted to attempt alphanumeric or algorithmic trace.
ONE_TIME_USE_ID	No PDS match found	alphanumeric trace live	FALSE	FALSE	FALSE	FALSE	1296	Not found in PDS, or in the MPS record bucket, so the Person_ID is a one-time-use ID. The last trace step attempted was alphanumeric trace, because there was not sufficient valid demographic information submitted to attempt algorithmic trace.
ONE_TIME_USE_ID	No PDS match found	algorithmic trace live	FALSE	FALSE	FALSE	FALSE	1951	Not found in PDS, or in the MPS record bucket, so the Person_ID is a one-time-use ID. All the previous tracing steps were attempted because sufficient valid demographic information was submitted.
ONE_TIME_USE_ID	No PDS match found	algorithmic trace live	FALSE	TRUE	FALSE	FALSE	205	Multiple valid matches found in PDS, so the Person_ID is a one-time-use ID because it was not possible to disambiguate the best match.
ONE_TIME_USE_ID	No PDS match found	No PDS tracing run	FALSE	FALSE	FALSE	FALSE	29890	Not found in PDS, or in the MPS record bucket, so the Person_ID is a one-time-use ID. No PDS tracing was run, because there was not sufficient valid demographic information to attempt any tracing steps.

6. Case studies

The examples in this section are based on real results observed in the processing of HES records through MPS, <u>with personal identifiable information all replaced by consistent fictitious values, hence making it impossible to identify real individuals</u>.

In this chapter we revisit three case studies from <u>The Person_ID handbook for HES users.</u> These case studies were chosen to illustrate MPS outputs which could be confusing to analysts, and the complex logic required to decipher them. With MPS Diagnostics, these three case studies would have been much easier to understand.

In case study 11 it would have been useful to have the Superseded NHS number flag.

In case study 13 it would have been useful to have the Multiple PDS matches flag.

In case study 14 it would have been useful to have the Last attempted MPS step.

Case Study 11: Two records with superseded versus current NHS numbers

Given two HES records with the following values:

HES record	NHS number	Gender	DOB	Postcode
1	4444444444	2	2003-03-03	null
2	5555555555	2	2003-03-03	LS1 4AP

FICTICIOUS DATA

We would receive the following response fields following processing from MPS Response file:

HES record	Person_ID	Matched Algorithm Indicator	Matched Confidence Percentage	Date Of Birth Score Percentage	Gender Score Percentage	Postcode Score Percentage
1	4444444444	1	100	null	null	null
2	4444444444	1	100	0	0	0

FICTICIOUS DATA

Without the MPS Diagnostics metadata, our logic to decipher how this occurred was as follows:

For HES record 1, we can see a match is made in the cross-check trace step within DPS as indicated by the *MatchedAlgorithmIndicator* value of 1. This means that an exact match was possible using the *NHS number* and *DOB* provided by the PDS records cached within DPS.

HES record 2 has a different *NHS number* but is matched to the same *Person_ID* as HES record 1 and has the same *MatchedAlgorithmIndicator* value of 1. *DOB*, *gender* and *postcode* score percentage fields have values of 0, which means that record 2 was cross-check traced in Spine. The success of the tracing to a *Person_ID* associated with a different *NHS number* allows us to infer that 555555555 is an invalid *NHS number* which has been superseded by 4444444444. Cross-check trace in DPS does not return matches where *NHS number*s are superseded, while cross-check trace in Spine was able to recognize the match because it also checks for superseded *NHS number*s.

With the MPS Diagnostics metadata, the additional columns show more clearly what has happened:

HES record	Person_ID	MPS SUCCESSFUL STEP	SUPERSEDED NHS NUMBER FLAG	NHS NUMBER HISTORY LIST
1	4444444444	CCT cached	False	[5555555555555]
2	4444444444	CCT live	True	[5555555555555]

FICTICIOUS DATA

Case Study 13: One-time-use ID *generated as no matches at any stage*

Given a HES record with the following values:

HES record	NHS number	Gender	Date of Birth	Postcode
1	null	1	2003-03-03	LS1 4AP

FICTICIOUS DATA

The following response fields could be returned by MPS:

HES record	Person_ID	Matched Algorithm Indicator	Matched Confidence Percentage	Date Of Birth Score Percentage	Gender Score Percentage	Postcode Score Percentage
1	U123KE3ABC	4	0	0	0	0

FICTICIOUS DATA

Without the MPS Diagnostics metadata, our logic to decipher how this occurred was as follows:

The *MatchedAlgorithmIndicator* value of 4 with *MatchedConfidencePercentage* of 0 indicates that none of the trace steps returned a match against the PDS records or the MPS record bucket. This could indicate that there are no individuals which match the identifying characteristics provided, or that multiple matches were returned and MPS was unable to determine a single match. This record has valid *DOB* and *postcode*, so it contains sufficient information to create a new *MPS_ID*. However, the *Person_ID* begins with 'U' indicating that a *one-time-use-ID* was generated for this record, hence we conclude that multiple NHS numbers were matched, and algorithmic trace could not resolve the match.

With the MPS Diagnostics metadata, the additional columns show more clearly what has happened:

HES record	Person_ID	MPS SUCCESSFUL STEP	MPS LAST STEP ATTEMPTED	MULTIPLE PDS MATCHES FLAG
1	U123KE3ABC	No PDS match found	algorithmic trace live	True

FICTICIOUS DATA

Case Study 14: Invalid DOB *results in no matches at any stage*

Given a HES record with the following values:

HES record	NHS number	Gender	Date of Birth	Postcode
1	3333333333	2	1800-01-01	LS1 4AP

FICTICIOUS DATA

The following response fields could be returned by MPS:

HES record	Person_ID	Matched Algorithm Indicator	Matched Confidence Percentage	Date Of Birth Score Percentage	Gender Score Percentage	Postcode Score Percentage
1	U123KE3ABC	0	0	0	0	0

FICTICIOUS DATA

Without the MPS Diagnostics metadata, our logic to decipher how this occurred was as follows:

The *MatchedAlgorithmIndicator* value of 0 indicates that all steps were skipped. This is because every trace step requires *DOB*, but in this case the *DOB* was recognised as invalid.

The *Person_ID* begins with 'U' indicating that a *one-time-use-ID* was generated for this record, hence there were no sufficient information to even generate a new *MPS_ID*.

With the MPS Diagnostics metadata, the additional columns show more clearly what has happened:

HES record	Person_ID	MPS SUCCESSFUL STEP	MPS LAST STEP ATTEMPTED
1	U123KE3ABC	No PDS tracing run	No PDS tracing run

FICTICIOUS DATA