NHS
Business Services Authority

personMatchR Package
Quality Review

Accuracy based on matching between NHS Electronic Prescribing data and Personal Demographic Service (PDS) data

Introduction

The personMatchR package (available on GitHub) provides functionality to allow two datasets to be compared to identify where individuals can be matched across both datasets, based on the individuals forename, surname, date of birth and postcode.

This document identifies the accuracy levels that have been identified for the output from this package.

The personMatchR package includes functionality to format and standardise an individual's forename, surname, date of birth and postcode, which enables a higher level of accuracy. Despite this, the accuracy achieved from the personMatchR package will strongly depend on the data being used as inputs for the process. This document may help identify where records are likely to produce a match or not.

Summary

To test the accuracy of the personMatchR package, data from NHS Electronic Prescribing (EPS) was compared with data from the Personal Demographic Service (PDS). Both these datasets include the patient's NHS number as an identifier, which could be used to confirm correct matches between the two datasets.

The package has the capacity to identify two match types, 'exact' and 'confident' matches. An exact match is where an individual shares the exact same forename, surname, date of birth and postcode across two datasets. A confident match is where these four fields are not all identical, yet similar enough to be confident that the same individual is being matched across datasets. If two records across datasets shared the exact same forename, surname, and postcode, along with a sufficiently similar date of birth, this would be one example of a confident match.

The ability to identify confident matches across datasets was a key feature of the package. However, confident matches have the potential to be incorrect or be duplicate matches. Processing of the matched output was required to deal with issues such as duplicate matches.

Both EPS and PDS datasets may include multiple variations of an individual's personal information, such as when an individual moves house or changes their surname. This may provide multiple opportunities for matches to be identified for a single individual. The nature of the matching process also means it is possible for a single individual to be matched against multiple other individuals. As a result, some minor processing is required to identify individual matches between NHS numbers from the output produced by the personMatchR package.

In total, 99.9% of the NHS numbers identified in the EPS data were correctly matched to the same NHS number in the PDS data based on the matching performed using the patient's personal information.

The remaining 0.1% included some cases where individuals with different NHS numbers were matched against each other, either on their own or in addition to "correct" matches. It should be noted that in rare circumstances a patient can be issued multiple NHS numbers and therefore different NHS numbers could still belong to the same individual.

In less the 0.05% of cases the NHS number in EPS could not be matched to any PDS records, despite PDS holding data for these NHS numbers. In these circumstances the root cause could generally be traced back to either low quality information in the datasets or the two datasets holding notably different information against the same NHS numbers.

Initial analysis suggests an excellent level of accuracy when the patient records exist in both datasets. There was however a risk that the package might identify an incorrect 'confident' match when no match should have been possible. This was shown not to be the case through a test dataset specifically checking for false positives.

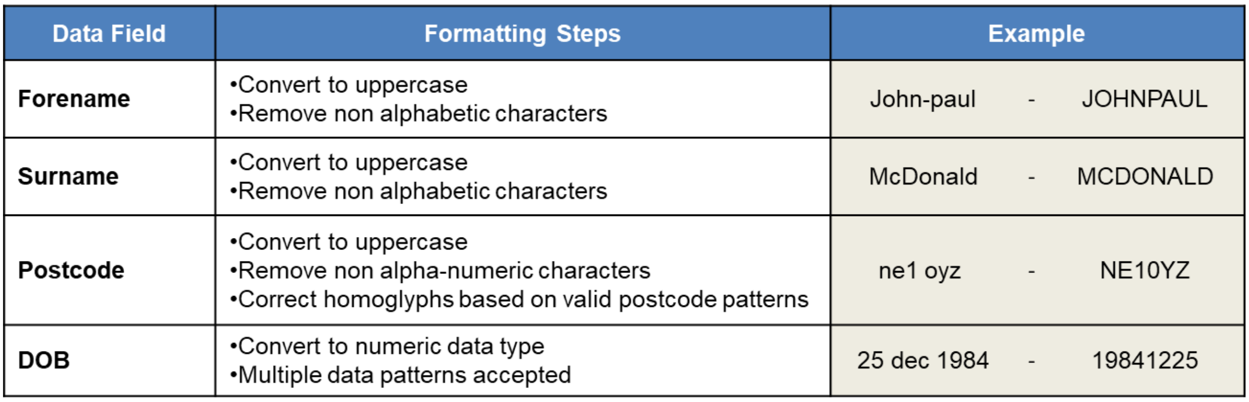
Data Overview

For this review, the matching process has been applied to data taken from electronic prescribing during the 2021/22 financial year, and data extracted from PDS.

Summary information for the PDS data and EPS data can be found as appendices at the end of this document

Data Formatting

Prior to the matching process, the formatting functions within the personMatchR have been applied to both sets of data. The 'Example' column shows how an individual's personal information required for matching might look before and after being processed.



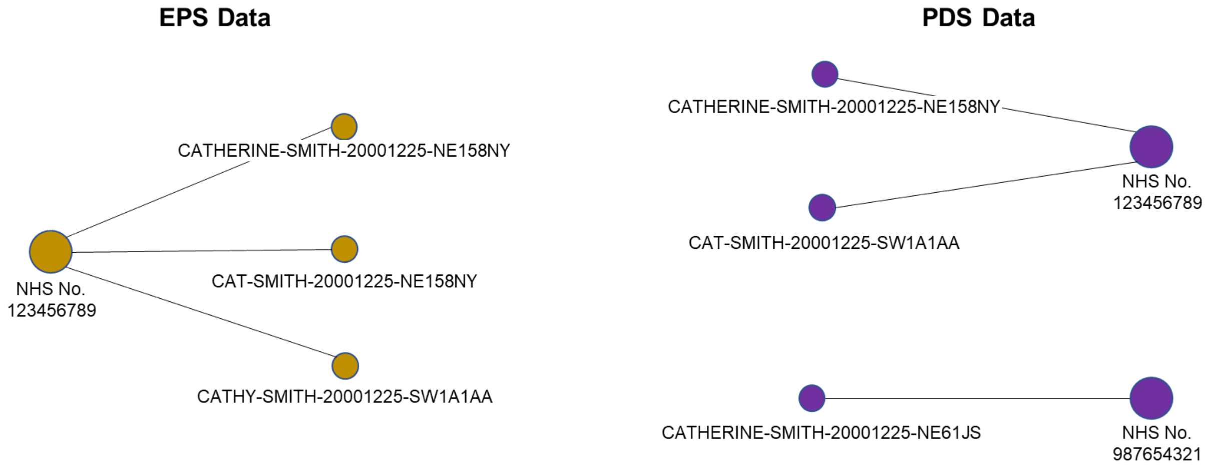
Data Field	Formatting Steps	Example
Forename	•Convert to uppercase •Remove non alphabetic characters	John-paul - JOHNPAUL
Surname	•Convert to uppercase •Remove non alphabetic characters	McDonald - MCDONALD
Postcode	•Convert to uppercase •Remove non alpha-numeric characters •Correct homoglyphs based on valid postcode patterns	ne1 oyz - NE10YZ
DOB	•Convert to numeric data type •Multiple data patterns accepted	25 dec 1984 - 19841225

Match Output Processing

The personMatchR package will simply perform matching across the two datasets that have been supplied as inputs, returning details of the identified matches. An individual may appear multiple times in a dataset, and each instance in turn has the potential to be matched against multiple other individuals.

The match output should not typically be immediately used within for analysis, as it will often require some processing. The extent and nature of the processing will depend on the use case.

For this quality review, as we are supplying data with multiple variations of personal information for each patient, some manipulation is required to get the data to report at an NHS number level.


EPS Data
CATHERINE-SMITH-20001225-NE158NY
NHS No.
123456789
CAT-SMITH-20001225-NE158NY
CATHY-SMITH-20001225-SW1A1AA
PDS Data
CATHERINE-SMITH-20001225-NE158NY
NHS No.
123456789
CAT-SMITH-20001225-SW1A1AA
CATHERINE-SMITH-20001225-NE61JS
NHS No.
987654321

The image above highlights the fact that different variations of personal information could match across both the EPS and PDS datasets. Some basic manipulation is required to simply identify which NHS number in EPS has matched to which NHS number in the PDS data. Where multiple variations of personal information have matched the same two NHS numbers, only the best match will be retained.

Following this manipulation, matches will only be passed for analysis/review where:

- An NHS number in EPS was only confidently matched to a single NHS number in PDS
- The NHS number in EPS matched exactly across the four pieces of personal information (forename, surname, date of birth, postcode).

Matches will not be included for analysis/review where:

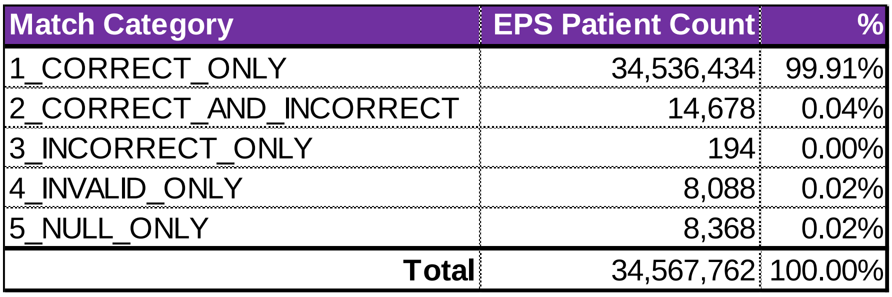
- The same NHS number in EPS matches (‘confident’ match) to multiple NHS numbers in PDS.
- The same NHS number in PDS would be assigned to multiple NHS numbers in EPS.

Match Results

Overall Result Summary

For over 99.9% of records, the personMatchR package was able to find the correct match, identified by records in both datasets having the same NHS number, without any other matches being identified.

The remaining records included some false positives and some cases where no match was identified.



Match Category	EPS Patient Count	%
1_CORRECT_ONLY	34,536,434	99.91%
2_CORRECT_AND_INCORRECT	14,678	0.04%
3_INCORRECT_ONLY	194	0.00%
4_INVALID_ONLY	8,088	0.02%
5_NULL_ONLY	8,368	0.02%
Total	34,567,762	100.00%

Correct Only Matches

These records represent the cases where the personMatchR package has identified the correct match from the dataset and represents 99.9% of cases.

Correct and Incorrect Matches Identified

In 0.04% of cases the NHS number from EPS matched both to the correct NHS number in PDS and an incorrect NHS number. However, by incorrect we simply mean different to the NHS number in EPS.

In all these cases the matches identified were all exact matches on all four pieces of personal information compared.

Within the PDS dataset it is known that some patients can receive more than one NHS number, and this is the most likely explanation for this outcome, where the EPS record has simply matched to all the numbers that exist for a patient and is likely to be the preferable outcome in this scenario.

The only alternative would be having two people living at the same postcode with the exact same name and date of birth, which would be impossible to identify without directly contacting individuals. Looking at a random sample of these matches and identifying the full address information confirmed that in all cases checked, the full address and not just the postcode matched for both NHS numbers.

Incorrect Only Matches

In less than 0.001% of cases the personMatchR package only produced a confident match to a different NHS number, although it must be noted that some individuals can potentially have multiple NHS numbers and therefore although the NHS numbers do not match, they may still be the same person.

When we looked at the “incorrect” NHS numbers that had been matched, 92% of these matched exactly across all pieces of personal information which would strongly suggest that this was the same patient with a different NHS number.

When reviewing why the “correct” NHS number in PDS had not been matched against typically this was because the four pieces of personal information could not be exactly matched between EPS and PDS. With no exact matches available the process will then look at the “confident” matches, but where these confident matches include matches to multiple potential options these all need to be ignored as there is no way to determine the “correct” choice from the potential options. Therefore, only the exact matches had been taken as viable options.

Invalid Only Matches

In 0.02% of cases, the NHS number from EPS did not produce any valid matches as all potential matches were flagged as invalid.

The matches were flagged as invalid for one of two reasons:

- multiple potential “confident” matches with no viable method to separate these without relying on a unique patient identifier such as NHS number (which were ignored for the purpose of this testing).
- the same record from PDS would have been matched to multiple records in EPS resulting in duplication. This could be the cases where patients have multiple NHS numbers and therefore match to multiple different NHS numbers.

For 94% of the Invalid Only Matches, the NHS number in EPS was being matched to both the “correct” and one or more other NHS numbers in PDS, and therefore were marked invalid as it would not be possible to determine the “correct” match from the potential options.

Only Null Matches
In 0.02% of cases the NHS number from EPS could not be matched to any of the records in PDS, despite the PDS data holding a record for the NHS number. This would be a contender for a false negative outcome as we would expect a match to be available.

One potential cause for this would be where the NHS numbers are missing some personal information in either of the datasets that may limit the ability of the personMatchR package to find a match. However, upon review over 99% of these NHS numbers had at least one record in both EPS and PDS with all personal information available.

When performing a manual review of these records one thing that appeared commonly was the EPS forename to be listed as “Baby”. By contrast, the PDS record would have the registered forename, which would explain why matches were not being found. Of the 8.4 thousand NHS numbers in EPS that did not get any match, 37% of these had recorded the forename “Baby”.

A random sample of 100 records, excluding those with “Baby” as the forename, was taken so the personal information could be manually compared between both datasets.
10 had slightly different information for forename, surname and postcode.
13 had notably different forenames as the only difference between datasets
The rest were different across a couple of fields, typically the postcode and one of the name fields.

The sample provides confidence that these records were not matched simply because the personal information data in EPS and PDS were notably different preventing a match from being possible without having a key identifier such as NHS number.

Potential false positive matches

The initial test case shows the personMatchR package has excellent accuracy when both datasets have the same people in them and therefore the correct matches are possible.

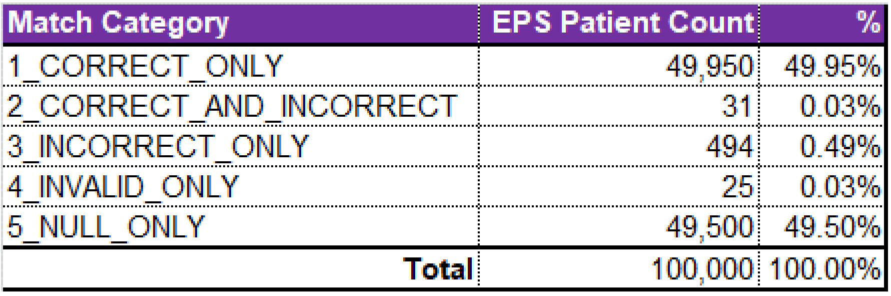
In many cases though, one of the datasets may not include the “correct” records to match to and we need to be confident that the personMatchR package will not then produce false positive matches by matching against the “best” possible record, unless this has a high chance of being correct.

To test the frequency of incorrect ‘confident’ matches where an exact match was not possible, a second test case was produced using the original EPS and PDS datasets:

- A random sample of 100 thousand patients was taken from EPS and all EPS data for these patients retained.
- Half of these 100 thousand patients were removed (based on matching NHS numbers) from the PDS dataset.

If the personMatchR model is not providing incorrect ‘confident’ matches where no matching should now be possible, we would expect half of the 100 thousand patients to be matched correctly, where their data exists in PDS and roughly half to not produce a match as there corresponding records have been removed from the PDS dataset.

As identified in the table below, this test case produced the desired outcome, showing that the package is not providing incorrect ‘confident’ matches when no match should be possible.



Match Category	EPS Patient Count	%
1_CORRECT_ONLY	49,950	49.95%
2_CORRECT_AND_INCORRECT	31	0.03%
3_INCORRECT_ONLY	494	0.49%
4_INVALID_ONLY	25	0.03%
5_NULL_ONLY	49,500	49.50%
Total	100,000	100.00%

APPENDICES

PDS Data

The PDS data represents the patient information supplied based on requests from NHSBSA for patient data for NHS numbers that have appeared in the NHS Prescription dataset. The data from PDS will include the patient’s name, address, and date of birth.

NHSBSA only request data extracts from PDS on a periodic basis, with the response from PDS based on the patient information at the point in time the response was produced. Therefore, the patient data returned by PDS may not match the patient information at the point of prescribing activity.

For this quality review, all historic responses from PDS have been included to give the biggest range of personal information for each NHS number. For any individual NHS number, multiple variations of personal information could exist where the patient’s information has changed over time (e.g. change of address). Additionally, PDS responses have only been included where the patient could be matched, and personal information supplied.

The dataset includes 83.0 million records, covering 56.3 million distinct NHS numbers, with each NHS number having between 1 and 31 distinct combinations of personal information (forename, surname, date of birth and postcode). Two-thirds (67.2%) of the NHS numbers only have a single combination of personal information, with a further 22.8% of NHS numbers having just two combinations of information. Only 0.9% of NHS numbers can be identified against five or more combinations of personal information.

The overwhelming majority (99.8%) of NHS numbers have at least one record where all pieces of the personal information are available. Where NHS numbers do not have any records with all pieces of information available, typically this is because the postcode is missing.

EPS Data

The EPS data represents the patient information captured from the NHS Electronic Prescribing System, representing the patient information supplied at the moment of prescribing. The data from EPS will include the patient's name, address, and date of birth.

Data has been limited to data relating to NHS prescribing taking place in 2021/22. Patient information captured in EPS could change from prescription to prescription if the patient's data has been updated on GP prescribing systems.

For this quality review data has been restricted to NHS numbers where the NHS number can be identified within the response data from PDS, to ensure the match process will have a "correct" record to match against.

The dataset includes 36.8 million records, covering 34.6 million distinct NHS numbers, with each NHS number having between 1 and 29 distinct combinations of personal information (forename, surname, date of birth and postcode). Most (94.0%) of the NHS numbers only have a single combination of personal information, with a further 5.6% of NHS numbers having just two combinations of information. Only 0.4% of NHS numbers can be identified against three or more combinations of personal information.

The overwhelming majority (99.97%) of NHS numbers have at least one record where all pieces of the personal information are available. Where NHS numbers do not have any records with all pieces of information available, typically this is because the postcode is missing.