# Match and Merge (Entity Resolution)

West Midlands Police / Data Analytics Lab

March 2025

# 1 Contents

## 2 Executive Summary

This briefing paper outlines the requirement for a match and merge process of people within (and across) WMP systems (a process known in other sectors as entity resolution).

The match and merge process was developed with the intention of identifying duplicate records that related to the same individual, but which might not have been identified as such due to small differences in the details recorded. Examples include where a name is submitted without a date of birth, and thus the record is not linked to a pre-existing record where the date of birth was captured. Another common example would be where the name of the individual is spelt differently in two different records, or where there is a typographical error relating to the date of birth.

As much as there is risk inherent in failing to identify duplicate records; there is equally risk in incorrectly merging records relating to different individuals

Due to data entry errors, untruths, etc. exact matching on names, etc. will likely under-merge. For this reason fuzzy matching by way of the Levenshtein distance is used.

In late 2024, Information Management raised concerns that there were a number of records identified as matches that were incorrect.

For this reason a Gold group has been convened to assess the most appropriate way forward for a match and merge process.

# 3 Introduction

This briefing paper outlines the requirement for a match and merge process of people within (and across) WMP systems (a process known in other sectors as entity resolution).

The match and merge process was developed with the intention of identifying duplicate records that related to the same individual, but which might not have been identified as such due to small differences in the details recorded. Examples include where a name is submitted without a date of birth, and thus the record is not linked to a pre-existing record where the date of birth was captured. Another common example would be where the name of the individual is spelt differently in two different records, or where there is a typographical error relating to the date of birth.

Policing relies on information held in its systems for a range of purposes, such as informing intelligence-based risk assessments; or making disposal decisions at the conclusion of an investigation. The police also rely on these systems for the purposes of sharing information with partners or making vetting and barring decisions.

Because of these different uses, failing to match duplicate records can result in critical information being missed that in turn leads to inappropriate decisions being made by police and partners. The Soham murder case is a sombre illustration of the potential consequences of missing crucial information.

It is for precisely these reasons that the report following the Joint Targeted Area Inspection (JTAI) of Solihull in 2022 identified the following improvement need:

'West Midlands Police need to take urgent action to improve the quality of information held on the 'Connect' system to make sure that links to connected individuals are present and accurate, and to reduce multiple records held against the same person, so that risk to children can be clearly seen, recognised, and shared when appropriate. Inspectors are concerned about incomplete records within the police 'Connect' system. Inspectors saw examples of separate records for the same person (because a name had been spelled incorrectly), children not linked on the system to their parents/carers, siblings or significant others and connections between children and those who pose a risk. This means that when officers and staff research 'Connect', they may miss important information, potentially leaving children at risk of significant harm.'

In response to this issue, changes were made to the Connect system at a significant cost that were intended to ensure certain duplicate records would be automatically identified and merged. The rules governing this automated process are described below in the table labelled 'phase 1'. Despite this process, it was noted that in August 2022, there were an estimated 250,000 records that required merging. It was considered that the methodology employed by the automated process would still leave a significant number of those records unmerged, and so the improvement plan identified a further necessary step:

*Commission WMP data lab to develop a match list of Person duplicates.*

Again, the process developed by the Data Analytics Lab (DAL) will be described in further detail below, but it is sufficient to note at this juncture that it expanded the number of records that could be considered a match beyond those identified by Connect[1].

Each week, DAL would run a search using the matching code and produce a list of results that were passed to IT&D. IT&D would then use a robot known as 'Radical' to edit Connect such that the duplicate records identified would be consolidated under a new 'Golden ID' (a unique reference relating to the newly merged records). From the outset, it was acknowledged that some of the records identified for matching might be erroneous. Even with an optimal process, this can occur for a variety of reasons, such as where an individual deliberately gives false details; or where two distinct individuals have very similar details (twins who live together for example, would share a date of birth, surname and address). As subsequent modelling has revealed, it can also occur simply because of the probability of unique individuals sharing the same personal details (see below for further explanation).

For these reasons, the match-and-merge processes were supposed to be subject to the failsafe of a schedule of manual dip-samples by staff in Information Management, with incorrectly matched logs separated and flagged to prevent future merging.

The combined Connect and DAL processes have been running since 2023, and approximately 332,200 records have been merged.

In late 2024, Information Management raised concerns that there were a number of records identified as matches that were incorrect. One notable example was a male and female who shared a common surname, a similar first name and the same PNC ID, but who were nonetheless completely unique individuals (note that this example arose because of the same PNC ID rather than the DAL process).

As much as there is risk inherent in failing to identify duplicate records; there is equally risk in incorrectly merging records relating to different individuals; this is because decisions could be made which have an adverse impact on an individual because of information that is falsely attributed to them. For example, a vetting or barring decision could result in the loss of potential employment; or an individual could be subject to a more significant sanction following investigation due to a mistaken assessment of antecedents. In an extreme scenario, a firearms intelligence assessment might incorrectly conclude a threat that results in disproportionate tactical responses.

It is on the basis of the discovery of these incorrectly merged records that Op Brightmind was established in late 2024.

[1] The process was originally developed to match nominals across different systems (e.g. Crimes and ICIS or now Connect and Compact). This requirement still exists due to the types of analyses that the Lab undertakes.

# 4 Matching Errors

The rules for the automated Connect merging process and the DAL/Radical process are different. Initial scoping would suggest that all of the incorrectly merged records that have been identified thus far have occurred as a result of the DAL/Radical process, rather than the Connect process.

The DAL process considers five key data points for the purposes of a identifying a potential match:

- Forename
- Surname
- Date of birth
- PNC ID
- CRO ID

Originally, the rules were configured to require a match of names, DOB and PNC/CRO ID; this was relaxed after dip-sampling showed too many potential duplicates were missed, mainly because circa 99% of all records lack PNC ID and CRO IDs. Subsequently, a match of surname, first name and either date of birth or PNC ID/CRO ID is considered sufficient to propose a merger

A match is defined either as 'exact' or 'fuzzy'. An exact match is as the term suggests, with no difference between spelling/format of DOB/PNC ID etc. A fuzzy match, by contrast, is one that is different by no more than a Levenshtein Distance of two characters. The Levenshtein distance between words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. In this case, therefore, the name or date of birth can have two different characters and still be considered a match.

The use of Levenshtein distance is designed to deal with typographical or spelling errors (e.g. where 'Katherine' is spelt as 'Catherine'; or where a date of birth is erroneously recorded as 31/01/1970, as opposed to 13/01/1970.

The choice to allow a distance of two was an attempt to strike the balance between the risk of incorrect mergers and missing necessary mergers. It means that the surname Hawkins, for example, could be matched with Howkins but Steve could not be matched with Stephen.

The logic of the rule makes sense when comparing two isolated names. What has emerged from analysis of the erroneous merges is that the coding of the algorithm used to propose matches did not take account of the merging of records over time. Within a Golden ID, there might be a record with the original, correct spelling of a name; and then a subsequent merged record with a name that was within two Levenshtein distance of the correct spelling of the original name. Over time, however, a third name could be proposed that fell within the acceptable Levenshtein distance of the merged spelling, yet be more than the acceptable distance of the original name. Taking the example above, John Hawkins born 13/1/70 might be (correctly) matched with John Howkins born 13/1/70. However, if John Tomkins born 31/1/70 is then arrested, a merge is proposed because although the Levenshtein distances for DOB and name are not close enough to match with Hawkins, they are close enough to match with Howkins.

In this example, the match is proposed even though the PNC ID number is either missing, or different, because the fuzzy match of a name and fuzzy match of the DOB are acceptable differences. The result is an incorrectly merged record.

Furthermore, an assumption was made that PNC ID and CRO IDs were reliable. In the case referred to at the outset of this paper, a male and female were merged because of a fuzzy match on first name, plus an exact match for surname and PNC ID. However, the PNC ID was incorrectly applied to one of the individuals by the PNC Bureau: despite the significant difference in DOB and obvious difference in gender, this meant that the record was merged.

# 5 Gold Strategy and Key Decisions

The Op. Brightmind Gold has been convened by ACC Welsted with two key aims:

1) To attempt to identify the volume of over-merged records

2) To revise current processes to reduce the risk of further erroneous matches

## 5.1 Volumes of over-merged records

Whilst a number of examples of erroneous matches have been identified, it is not currently possible to quantify how many such cases there are. It is assumed that a significant proportion of records that have been merged will have been suitable for merging; however, there is no way of differentiating those from the erroneous mergers without a manual check.

In a bid to target the manual efforts most efficiently, the DAL have produced a spreadsheet which contains information of Golden IDs (GIDs) with multiple names, dates of birth and PNC ID reference numbers contained. These lists can be filtered to identify, for example, the number of GIDs with more than two distinct names, dates of birth or PNC IDs. The more distinct names/dates of birth etc. a GID contains, the more likely it is to have merged records that exceed the Levenshtein distance rules; and therefore, the more likely it is to have merged records incorrectly. Manual dip-samples can prioritise the GIDs with the highest number of distinct names etc. and where errors are identified, de-merge records. This piece of work is ongoing at the time of writing.

## 5.2 Merging Process Revisions

From enquiries with peer forces and other organisations, there appear to be some common principles applied, but no definitive approach to match and merge. This is essentially because there has been no approach identified that can perfectly differentiate records than need to be merged from those that should not. **Rather, the approach taken will have to reflect an organisational 'risk appetite'.**

In light of this ambiguity, one of the key decisions from Gold was to pause all matching and merging processes (both the automated Connect process, and the ancillary DAL/Radical process) pending agreement of a revised process.

In determining the appropriate risk appetite, it is important to recognise that there are essentially two 'competing' risks (merging erroneously versus failing to merge records relating to the same individual). The more 'expansive' the criteria for merging, the lesser the risk of failing to merge, but the greater the risk of merging erroneously. Conversely, the more restrictive the rules are, the greater the risk of missing a record that should be merged; but the lesser the risk of merging incorrectly. The decision, therefore, is where on a continuum the merge rules should sit:

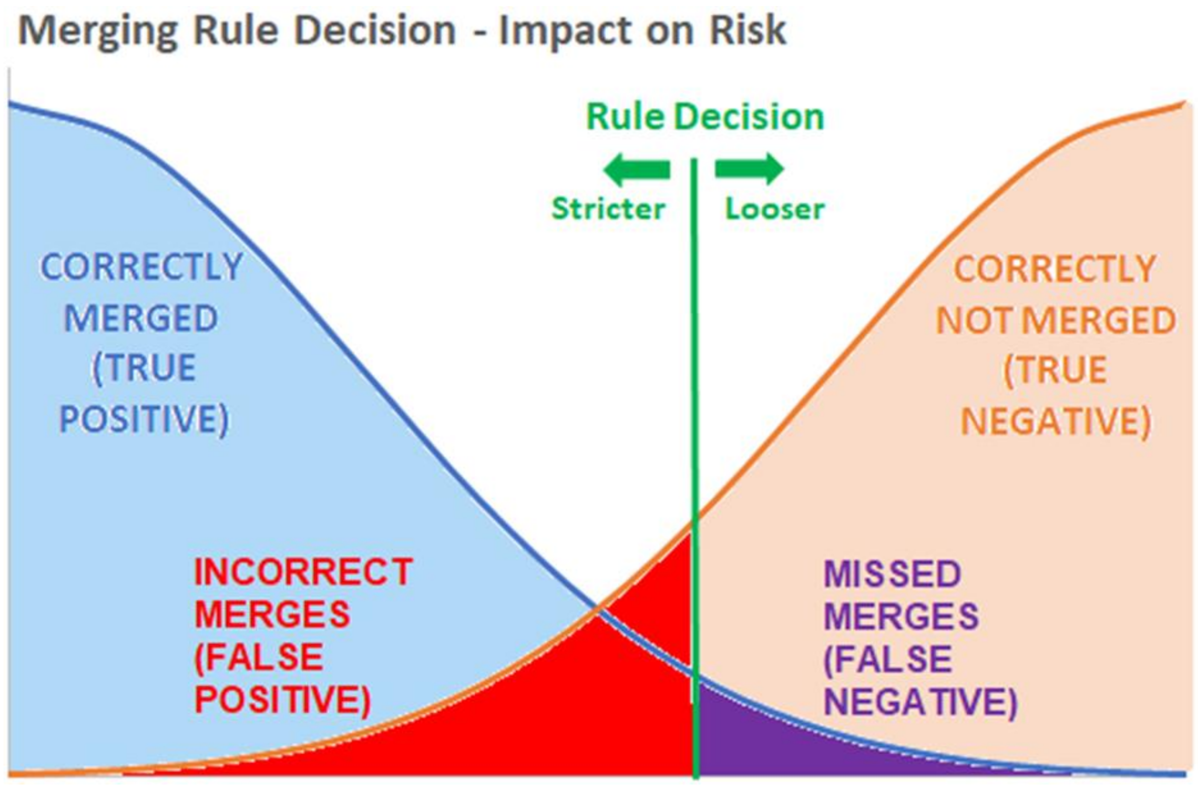

Merging Rule Decision - Impact on Risk
Rule Decision
Stricter
Looser
CORRECTLY MERGED (TRUE POSITIVE)
CORRECTLY NOT MERGED (TRUE NEGATIVE)
INCORRECT MERGES (FALSE POSITIVE)
MISSED MERGES (FALSE NEGATIVE)

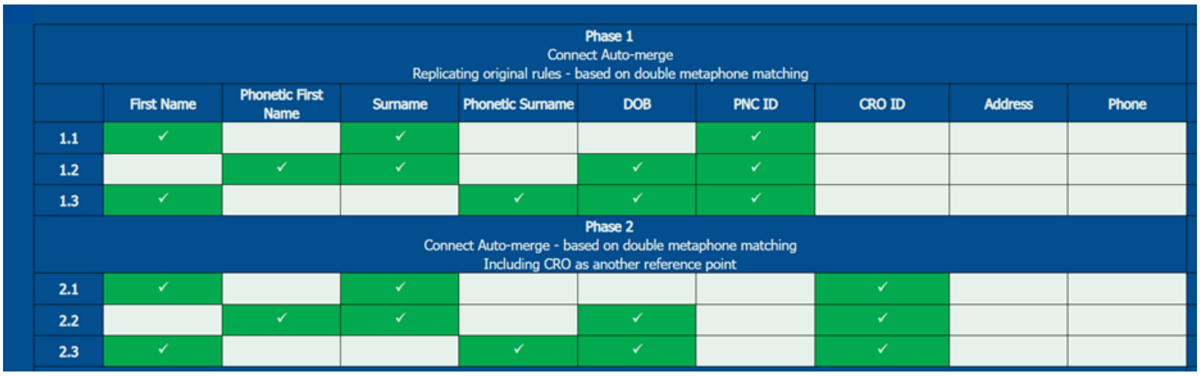
*Figure 1: There is error probabilities on both sides*

It is also important to note that the DAL carried out some modelling to ascertain the likelihood of two distinct individuals sharing the same first name, surname and date of birth purely at random. The details of these calculations are included in the appendix, but for the purposes of this paper, it is relevant to note that the probability of such an occurrence is so high as to be considered certain to occur. It is important to note that this does not mean it is certain that there would be erroneous merges: rather it means that the potential for such an erroneous merge exists even if the rules for a match were set in their most restrictive form (i.e. only exact matches of names and dates of birth would be merged).

In the absence of any authoritative guidance on what merging rules should be set, a key decision from the Op Brightmind Gold is that a 'strict' set of rules will be applied as a baseline, with 'looser' rules being introduced through a series of phases, informed by dip-sampling of cases to ascertain the frequency of erroneous merges. It is important to recognise that such an iterative approach will not provide a counter-factual (i.e. all of the missed duplicates).

With this in mind, the first phase will see the resumption of the automated approach being used within the Connect system. This involves exact matches for at least three data points where the data points are first name, surname and PNC/CRO ID; or three exact matches for either first or surname, plus DOB and CRO/PNC ID, supplemented by a phonetic match for the name that was not an exact match.

*NB, Connect uses the Double Metaphone software. Double Metaphone identifies words that are spelt differently, but produce the same sound (e.g. John and Jon, or Katherine and Catherine).*

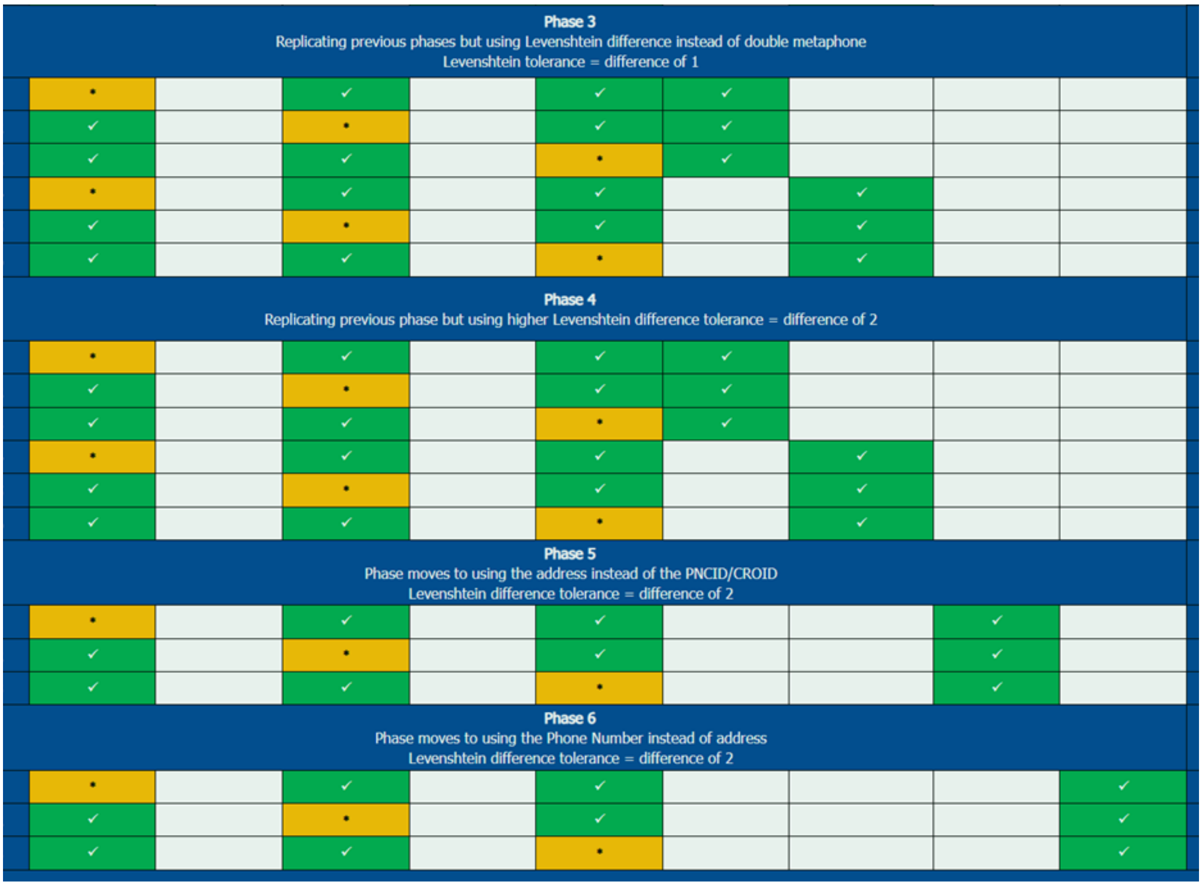


| | First Name | Phonetic First Name | Surname | Phonetic Surname | DOB | PNC ID | CRO ID | Address | Phone |
|---|---|---|---|---|---|---|---|---|---|
| Phase 1<br>Connect Auto-merge<br>Replicating original rules - based on double metaphone matching | | | | | | | | | |
| 1.1 | ✓ | | ✓ | | | ✓ | | | |
| 1.2 | | ✓ | ✓ | | ✓ | ✓ | | | |
| 1.3 | ✓ | | | ✓ | ✓ | ✓ | | | |
| Phase 2<br>Connect Auto-merge - based on double metaphone matching<br>Including CRO as another reference point | | | | | | | | | |
| 2.1 | ✓ | | ✓ | | | | ✓ | | |
| 2.2 | | ✓ | ✓ | | ✓ | | ✓ | | |
| 2.3 | ✓ | | | ✓ | ✓ | | ✓ | | |

*Figure 2: Phase 1 and 2 for Connect*

These rules are considered 'low-risk' in relation to erroneous matches, but will likely be higher risk when it comes to duplicate records that lack PNC/CRO ID. Circa 99% of all Connect records lack these data-points, meaning that merge rates will be very low.

The matching algorithm has been amended to reflect and testing of the phases below:



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Phase 3<br>Replicating previous phases but using Levenshtein difference instead of double metaphone<br>Levenshtein tolerance = difference of 1 | | | | | | | | | |
| | • | | ✓ | | ✓ | ✓ | | | |
| | ✓ | | • | | ✓ | ✓ | | | |
| | ✓ | | ✓ | | • | ✓ | | | |
| | • | | ✓ | | ✓ | | ✓ | | |
| | ✓ | | • | | ✓ | | ✓ | | |
| | ✓ | | ✓ | | • | | ✓ | | |
| Phase 4<br>Replicating previous phase but using higher Levenshtein difference tolerance = difference of 2 | | | | | | | | | |
| | • | | ✓ | | ✓ | ✓ | | | |
| | ✓ | | • | | ✓ | ✓ | | | |
| | ✓ | | ✓ | | • | ✓ | | | |
| | • | | ✓ | | ✓ | | ✓ | | |
| | ✓ | | • | | ✓ | | ✓ | | |
| | ✓ | | ✓ | | • | | ✓ | | |
| Phase 5<br>Phase moves to using the address instead of the PNCID/CROID<br>Levenshtein difference tolerance = difference of 2 | | | | | | | | | |
| | • | | ✓ | | ✓ | | | ✓ | |
| | ✓ | | • | | ✓ | | | ✓ | |
| | ✓ | | ✓ | | • | | | ✓ | |
| Phase 6<br>Phase moves to using the Phone Number instead of address<br>Levenshtein difference tolerance = difference of 2 | | | | | | | | | |
| | • | | ✓ | | ✓ | | | | ✓ |
| | ✓ | | • | | ✓ | | | | ✓ |
| | ✓ | | ✓ | | • | | | | ✓ |

*Figure 3: Phases 3 - 6 for match and merge process*

The initial phases will replicate the rules for phases 1 and 2 based on a Levenshtein distance of one then two, and introducing the additional data points of address and telephone number. Addresses are recorded with far greater frequency than PNC/CRO ID. However, the proportion of records where an address is recorded is circa 12% and those with matches to previous records for the same person are estimated to equal circa 18% of records. As such, there will still be a high number of duplicate records that will not be merged, even though they relate to the same individual. A similar issue exists in relation to telephone numbers.

*NB addresses will be matched on the basis of the unique location reference recorded within Connect, rather than the spelling/format of address as entered.*

## 5.3 Assessing Error

In order to have an idea of the potential effect of different configurations upon error rates (in the absence of samples) and provide an assessment of the size of samples needed for varying degrees of confidence simulations have been undertaken which incorporate errors into a matching process.

Missed Merges – Percentage of records that should have been merged, but have not be merged to the correct record (or any other record). These risk a nominal's whole history not to be joined together and missed, possibly leading to incorrect decisions in an investigation or in offender management.

Incorrect Merges – Percentage of records, that end up in a merged group that contains records that the record should not be merged with. This is the opposite of 'Missed' merges, with the risk of combining different nominals records together, again possibly leading to incorrect decisions in an investigation or in offender management.

Correct – Record either; should not be merged with anything and wasn't, or should be merged with another record, which was correctly identified by the merging process, with no additional incorrect records merged.

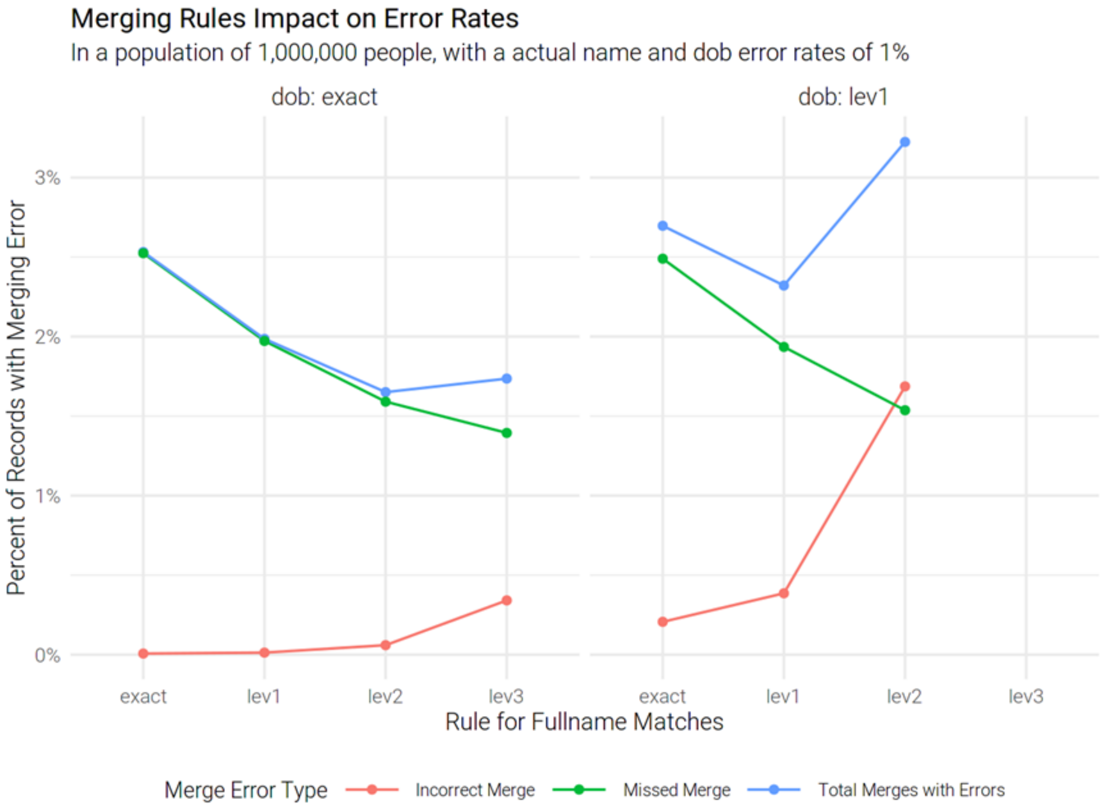


*Figure 4: The impact of potential rules on error rates*

Assessing the results, it can be seen that the combination of merge rules with the lowest total errors is an exact DOB match and a full name match with a Levenstein distance similarity of up to 2.

# 6 Conclusion

The Gold group process is currently working through the potential different configurations of matching rules to assess the potential effect with a view to adopting the 'best' approach to use.

It is recognised that there is no perfect approach to the process as there will be errors of both merging incorrectly and missing merges (and it is certain that within the West Midlands there will be people who exactly match names and dates of births purely by chance).

Any such process eventually agreed upon will therefore be subject to manual adjustments.

# Appendix

## Probability, Pairs and Proportions

Many forces (and the systems they use) around the country use some form of match and merge over nominals, often with varying approaches to fuzzy matching.

It may not be appreciated that there is some probability of people in a population having the same name and date of birth at random.

In a population of any appreciable size, the probability of different people (unknown to each other) having the same name and date of birth is 1, i.e. certainty. This is the probability of at least one match.

For example, the probability of the most commonly occurring name from an external dataset is:

1 in 3,846 (0.026%)

The probability of any particular day of the year is:

1 in 365 (0.27%)

and the probability of being born in any particular year (assuming there is a period of 70 years to cover) is:

1 in 70 (1.43%)

The overall probability of someone having this name and any birthday and year of birth is:

0.026% x 0.27% x 1.43% = 1 in 98,265,300

This probability is very low; however, the question of interest is 'how likely are we to find any matches given this probability'?

The probability of finding at least one match in a large population is 1 (certain) because of the number of pairs that are present in a population.

For example, despite the joint probability of any one person having a name and date of birth being low, a population of only 11,700 people is required for the probability of a match to be slightly greater than 50%; a population of 60,000 leads to a probability of 99.99999%.

*The population of the West Midlands force area is (as of the Census 2021) 2,919,650; with this number of people, the probability of a match is therefore certain.*

## High Probability, Low Proportions

**HOWEVER, the proportion of any population that do match the name and date of birth of someone else is low** (because this is in essence a different question).

For example, if we simulate a population of 2,900,000 different people with names taken from a list of (circa 6m) unique names (weighted according to occurrence) and the dates of birth using the probabilities noted above, then on average we find 0.03% match. This equates to 756 people who would be erroneously merged (378 pairs).

The starting point for the Connect system however is the person_object_ref level, which has circa 18.9 million records.

Using the above probabilities amongst the person_object_ref records it would therefore be certain that there will be at least one random match and in fact the probability of matching (1 in 3,816 or 0.03%) means there would likely be circa 9,880 records with the exact name and date of birth by random (excluding any effects associated with committing crime, etc.).