|  |  |  |  |
| --- | --- | --- | --- |
|  | Pairs | Dropping | Assumptions |
| Only first name different – minor spelling errors to completely different name | 412 | 13 | \* minor spelling errors in first name = matching pair  \* completely different first name, same gender or with unisex names, with same DOB, address = matching pair  https://genderchecker.com - this website used to confirm the gender. We are also assuming that there are no twins |
| Only last name different – minor spelling errors to completely different name | 283 | 0 | \* minor spelling errors in last name = matching pair  \* since every other element matches we can assume that it is the same person  even with completely different last name  \* we are also assuming that they are not twins  \* keeping all of them |
| Only month DOB element different | 65 | 0 | Only one element different, so assuming that they are the same person  \*keeping all of them |
| Only day DOB element different | 262 | 0 | Only one element different, so assuming that they are the same person  \*keeping all of them |
| Only year DOB element different | 856 | 0 | Some do have change in year 19 vs 20 example 1947 vs 2047 - this is definitely same person  Assumptions  Only one element different, so assuming that they are the same person  \*keeping all of them |
| First name and last name different | 74 | 8 | \*\*\* assumptions  \* we are assuming that there are no twins  \* first and last name interchanged - assume it is same person  \* if minor misspelling in both first and last name - assume it is same person  \* if first name or last name is completely different, and other element has minor  misspelling, and gender is same assume it is same person  \* female and very different last name and minor misspelling in first name, assume it  is same person  \* either first name and/or last name are completely different, but gender is different assume they  are different people  \* first name and last name are completely different, and gender is same assume they  are different people |
| First name and day of birth different | 50 | 35 | \* minor misspelling in first name and different day of birth ~ assume it is same  person  \* first name is very different, and confirming it is not middle name/different gender  on the complete aps/medstar dataset, assume it is different person  \* first name is very different, and confirming it is middle name on the  complete aps/medstar dataset, assume it is same person |
| First name and month of birth different | 6 | 1 | \* minor misspelling in first name and different month of birth ~ assume it is same  person  \* first name is very different, and confirming it is not middle name and different gender  on the complete aps/medstar dataset, assume it is different person  \* first name is very different,and confirming it is middle name on the  complete aps/medstar dataset, assume it is same person  \* first name is very different, same gender - assume it is same person |
| First name and year of birth different | 111 | 3 | \* minor misspelling in first name and different year of birth ~ assume it is same  person  \* first name is very different, and confirming it is not middle name and different gender  on the complete aps/medstar dataset, assume it is different person  \* first name is very different, and confirming it is middle name on the  complete aps/medstar dataset, assume it is same person  \*first name is very different, but year is 20xx vs 19xx, and same gender  assume it is same person (and not twins) |
| First name and address different | 355 | 7 | - very different first name, same address, same gender, - assume it is same person  - very different first name, same address, different gender - assume it is different person  - minor misspelling first name, same address, same gender - assume it is same person  - minor misspelling first name, different address, same gender - assume it is same person  - very different first name, different address, same gender -assume it is different person |
| Last name and day of birth different | 14 | 0 | Assume pair |
| Last name and month of birth different | 2 | 0 | Assume pair |
| Last name and year of birth different | 39 | 0 | Assume pair |
| Last name and address different | 250 | 29 | #reviewing those pairs with nonmatching last name and address  - minor typos in last name - assume same person  - very different last name - look at address -> minor typo in address - assume same person  - very different last name - look at address -> very different address --> confirm if there are different addresses in medstar/aps if available --> overlapping address --> assume same person  - very different last name - look at address -> very different address --> confirm if there are different addresses in medstar/aps if available --> non-overlapping address --> assume different person |
| Day of birth and Month of birth different | 25 | 0 | Assume all the observations are matching pairs due to   * Information collected from third person * Information collected from non-medical records * Typographical issue |
| Day of birth and Year of birth different | 35 | 0 | Assume all the observations are matching pairs due to   * Information collected from third person * Information collected from non-medical records * Typographical issue |
| Day of Birth and Address different | 62 | 0 | Assume all the observations are matching pairs due to   * Information collected from third person * Information collected from non-medical records * Typographical issue |
| Month of birth and year of birth different | 10 | 0 | Assume all the observations are matching pairs due to   * Information collected from third person * Information collected from non-medical records * Typographical issue |
| Month of birth and address different | 17 | 0 | Assume all the observations are matching pairs due to   * Information collected from third person * Information collected from non-medical records * Typographical issue |
| Year of birth and address different | 141 | 0 | * Information collected from third person * Information collected from non-medical records * Typographical issue |
| First name, last name and day of birth different | 0 | 0 |  |
| First name, last name and month of birth different | 0 | 0 |  |
| First name, last name and year of birth different | 0 | 0 |  |
| First name, last name and address different | 63 | 38 | ##First name, last name and address different  Assumptions  - if minor spelling error in first name, last name - assume same person  - if first and/or last name very different - address very different - assume different person  - if first and/or last name very different - address minor typo - same gender - assume different person  - if first and/or last name very different - address minor typo - different gender - assume different person  - if first name and last name are interchanged - assume same person regardless of address |
| First name, Day of birth and month of birth different | 83 | 55 | assumptions  - if first name very different - assume different person as they could be siblings/partners  - first name minor typo - assume same person |
| First name, Day of birth and Year of birth different | 71 | 63 | - if first name very different - assume different person as they could be siblings/partners  - first name minor typo - assume same person |
| First name, Day of birth and address different | 11 | 3 | - if first name very different - assume different person as they could be siblings/partners  - first name minor typo - assume same person |
| First name, month of birth and year of birth different | 53 | 48 | - if first name very different and/or different gender - assume different person as they could be siblings/partners  - first name minor typo - assume same person |
| First name, month of birth and address different | 6 | 2 | - if first name very different - assume different person as they could be siblings/partners  - first name minor typo - assume same person |
| First name, year of birth and address different | 15 | 0 | - if first name very different - assume different person as they could be siblings/partners  - first name minor typo - assume same person |
| Last name, day of birth and month of birth different | 12 | 12 | -Assume different person if very different last name as they could be living in a group home, apartment.  - Assume same person if minor typo in last name. |
| Last name, day of birth and year of birth different | 14 | 12 | - Assume different person if very different last name as they could be living in a group home, apartment.  - Assume same person if minor typo in last name. |
| Last name, day of birth and address different | 8 | 4 | - assume same person if minor typo in last name  - if last name completely different and address is different assume different person |
| Last name, month of birth and year of birth different | 2 | 2 | - assume same person if minor typo in last name  - if last name completely different, year is in 2000’s instead of 1900’s and month off by multiple months assume same person  - otherwise assume same person |
| Last name, month of birth and address different | 1 | 0 | - assume same person if minor typo in last name  - if last name completely different, month is off by multiple months and address different assume different person |
| Last name, year of birth and address different | 38 | 3 | - assume same person if minor typo in last name  - if last name completely different, year is off by multiple years and address different assume different person |
| Day of birth, month of birth and year of birth different | 96 | 0 | Assume same person as   * Very rare to have same first and last name and live in same address * Assume typo in all 3 elements of date of birth |
| Day of birth, month of birth and address different | 13 | 0 | * Assume same person if address has minor typo. * Else assume different person |
| Month of birth, year of birth and address different | 2 | 0 | * Assume same person if address has minor typo. * Else assume different person |
| Day of birth, year of birth and address different | 6 | 0 | * Assume same person if address has minor typo. * Else assume different person |
| First name, last name, day of birth, month of birth different | 0 | 0 |  |
| First name, last name, day of birth, year of birth different | 0 | 0 |  |
| First name, last name, day of birth, address different | 6 | 6 | * Assume same person if minor typo in first name, last name and/or address * If first name, last name completely different assume different person * If first name only different, minor typo in last name, minor typo in address, - same gender assume same person, different gender assume different person * -last name very different assume different person |
| First name, last name, month of birth, year of birth different | 0 |  |  |
| First name, last name, month of birth, address different | 0 |  |  |
| First name, last name, year of birth, address different | 0 |  |  |
| First name, day of birth, month of birth, year of birth different | 529 | 486 | * Minor typo in first name assume same person as it is rare that two people with almost similar first and last name reside at same address. * Major difference in first name and/or different gender assume different person as it could be partner, sibling or other relative |
| First name, day of birth, month of birth, address different | 3 | 3 | * Minor typo in first name assume same person as it is rare that two people with almost similar first and last name reside at same address. * Major difference in first name and/or different gender assume different person as it could be partner, sibling or other relative |
| First name, year of birth, month of birth, address different | 2 | 2 | * Minor typo in first name assume same person as it is rare that two people with almost similar first and last name reside at same address. * Major difference in first name and/or different gender assume different person as it could be partner, sibling or other relative |
| last name, day of birth, month of birth, year of birth different | 182 | 170 | * Minor typo in last name assume same person as it is rare that two people with almost similar first and last name reside at same address. * Major difference in last name assume different person as it could be partner, sibling or other relative or it could be some kind of group living facility or apartment complex |
| Last name, day of birth, month of birth, address different | 41 | 41 | - Minor typo in last name and minor typo in address assume same person  - Last name very different assume different person even if living in address where there is minor typo as we are unaware of type of living facility |
| Last name, year of birth, month of birth, address different | 11 | 11 | * Minor typo in last name and minor typo in address assume same person * Last name very different assume different person even if living in address where there is minor typo as we are unaware of type of living facility |
| Last name, year of birth, day of birth, address different | 23 | 23 | * Minor typo in last name and minor typo in address assume same person * Last name very different assume different person even if living in address where there is minor typo as we are unaware of type of living facility |
| Day of Birth, month of birth, year of birth and address different | 13 | 4 | * Minor typo in address, assume same person as it is rare for person with same first and last name to live in same address (with minor typo in address) * Different address assume different person |
| Other non-manually reviewed pair identified when creating dummy variable to identify pairs and non-pairs | 524 | 146 | Assumptions  - First name, last name - minor typo, DOB same assume same person regardless of address  - First name - very different, same last name, address and DOB assume same person  - Last name very different, same first name, address and DOB assume same person  - First name - very different, same last name, address and different DOB assume different person  - Last name very different, same first name, address and different DOB assume different person  - First name - very different, same last name, different address and same DOB assume different person  - Last name very different, same first name, different address and same DOB assume different person |

* # there are 27932 potential observations, 2206 non-pairs and 1048 observations that is still needed to be reviewed manually
* After reviewing final 1048 observations we ended up with 28688 observations (14344 pairs) and 2498 non-pairs
* Pairs where APS happened before Medstar – 9616 – so we will end up with 4728 pairs
* Pairs where APS happened a month after medstar -> 2923 -- so we end up with 1805 potential matches

Check conflicts on only – first name, last name

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| First name conflict |  | 16 |  |
| Last name conflict | 714 | 1 |  |

Resolving conflicts and choosing dataset

|  |  |  |
| --- | --- | --- |
|  | Instances with conflict | Dataset to use |
| First name | 593 | APS |
| Last name | 374 | APS |
| Full name | 1661 | APS |
| Birth month | 225 |  |
| Birth day | 336 |  |
| Birth year | 726 | Medstar |
| DOB | 965 | Medstar |
| Age | 3224 | Medstar |
| Street Address | 2484 | Medstar |
| Address Street Name | 2312 | Medstar |
| Address City | 906 | Medstar |
| Address Zip | 5198 | Medstar |
| Address number | 1800 | Medstar |

Medstar\_aps\_merged dataset dimensions – 51359 \* 112

After removal of unneeded variables dimesions 🡪 51359 \* 14

* Removing duplicate rows 🡪 dimensions 🡪 49569\*14