4-RecordLinkage

At the time of our study, the Neotree app required users to manually enter the automatically generated admission unique identifier (UID) into a free-text field when completing the outcome form. Therefore, the outcome UID is liable to typographical errors and is not a 100% reliable key to link admission and outcome forms. Thus, we linked records using the Fellegi-Sunter framework of probabilistic record linkage.

0.1 Create data frames for linkage

0.1.1 Linkage variables

There are 8 variables common to both admission and outcome forms:

1. UID
2. Birth weight
3. Gestation at birth
4. Occipitofrontal circumference at admission
5. Length at admission
6. Mode of delivery
7. Sex
8. Place of birth

These have the following levels of missingness in the admission forms:

```
## # A tibble: 8 x 3
##   variable               n_miss pct_miss
##   <chr>                   <int>    <dbl>
## 1 Admission.PlaceBirth     3637  87.9
## 2 Admission.BW               68   1.64
## 3 Admission.OFC               5   0.121
## 4 Admission.Gestation         1   0.0242
## 5 Admission.Length            1   0.0242
## 6 Admission.UID_alphanum      0   0
## 7 Admission.ModeDelivery      0   0
## 8 Admission.Gender            0   0
```

And in the outcome forms:

```
## # A tibble: 8 x 3
##   variable                     n_miss pct_miss
##   <chr>                         <int>    <dbl>
## 1 Discharge.GestBirth            2499  63.5
## 2 Discharge.OFCDis                248   6.30
## 3 Discharge.LengthDis             231   5.87
## 4 Discharge.BirthPlace             13   0.330
## 5 Discharge.Delivery                1   0.0254
```

```
## 6 Discharge.NeoTreeID_alphanum      0   0
## 7 Discharge.BWTDis                  0   0
## 8 Discharge.SexDis                  0   0
```

Note `Place of birth` has 87.9% missing values in the admission forms. It is also coded differently between admission and outcome forms:

```
## $admission
## # A tibble: 5 x 2
##   levels definition
##   <chr>  <chr>
## 1 BBA    born before arrival
## 2 HC     health centre
## 3 Home   home
## 4 Hosp   hospital
## 5 TBA    traditional birth attendant
##
## $outcome
## # A tibble: 4 x 2
##   levels definition
##   <chr>  <chr>
## 1 H      home
## 2 HCH    Harare Central Hospital
## 3 OtH    other clinic in Harare
## 4 OtR    other clinic outside Harare
```

Although `Gestation at birth` has 63.5% missing values in the outcome forms, it is a numeric variable and, therefore, coded the same between admission and outcome forms:

```
## $admission
## [1] "40" "41" "33" "40" "40" "38"
##
## $discharge
## [1] "34" "32" "31" "27" "34" "34"
```

Thus, we used 7 variables for record linkage (excluding `Place of birth`):

1. Unique ID*
2. Birth weight
3. Gestation at birth
4. Occipitofrontal circumference at admission
5. Length at admission
6. Mode of delivery
7. Sex
8. ~~Place of birth~~

* **Special note on UID:** After the first month of the project, healthcare workers were told to only enter the first 3 and last 3 characters of the UID in the outcome form. This is because the UID was initially long and laborious to type and, hence, prone to error. Therefore, most outcome form UIDs are 6 characters long, except those from the first month of the project. To avoid confusion, we used a substring of the full UIDs (called `uidsub`) for linkage. This substring consists of the first 3 and last 3 characters of the UID, converted to lowercase. E.g. (fictitious example),

```
##      full       uidsub
## [1,] "AB123456" "ab1456"
## [2,] "AB789012" "ab7012"
```

0.1.2 Linkage data frames

Below is a fictitious example of the data frame structure for record linkage:

```
## $admission
## # A tibble: 4 x 8
##   bw    gest  ofc   length mode  sex   session       uidsub
##   <chr> <chr> <chr> <chr>  <chr> <chr> <chr>         <chr>
## 1 3000  41    32    46     SVD   M     session 10000 ab1789
## 2 4000  40    37    46     ECS   F     session 10001 ab2567
## 3 1800  35    31    44     SVD   F     session 10002 ab3689
## 4 3500  40    33    48     SVD   M     session 10003 ab1478
##
## $outcome
## # A tibble: 4 x 8
##   bw    gest  ofc   length mode  sex   session        uidsub
##   <chr> <chr> <chr> <chr>  <chr> <chr> <chr>          <chr>
## 1 3320  <NA>  32    48     SVD   F     session 100000 cd3567
## 2 1900  32    32    47     ECS   F     session 100001 cd1378
## 3 1900  34    30    45     SVD   M     session 100002 cd8364
## 4 1300  32    29    39     ECS   M     session 100003 cd9246
```

0.2 Perform record linkage

0.2.1 Run linkage algorithm

We performed record linkage using the `fastLink` package by Enamorado, Fifield and Imai (https://github.com/kosukeimai/fastLink). Linkage is performed using the `fastLink::fastLink()` wrapper.

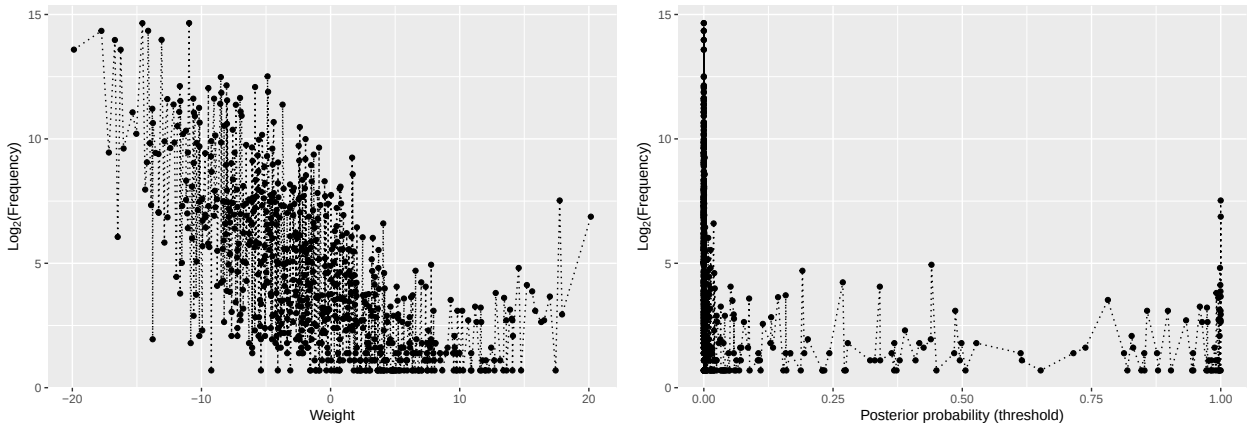
```
set.seed(123)

matches_out <- fastLink(
  dfA = adm_link,
  dfB = dis_link,
  varnames = c("uidsub", "bw", "gest", "ofc", "length", "mode", "sex"),
  stringdist.match = c("uidsub"), # use string dist matching on uidsub
  stringdist.method = "jw", # Jaro-Winkler
  jw.weight = .10, # Jaro-Winkler weight for prefix
  partial.match = c("uidsub"), # allow partial matching for uidsub
  cut.a = 0.96, # full string-distance match cut point (Winkler, 1990)
  cut.p = 0.88, # partial string-distance match cut point (Winkler, 1990)
  dedupe.matches = TRUE, # enforces one-to-one matching
  cond.indep = TRUE, # assuming conditional independence for Fellegi-Sunter model
  return.all = TRUE # sets threshold.match to 0.0001
)
```

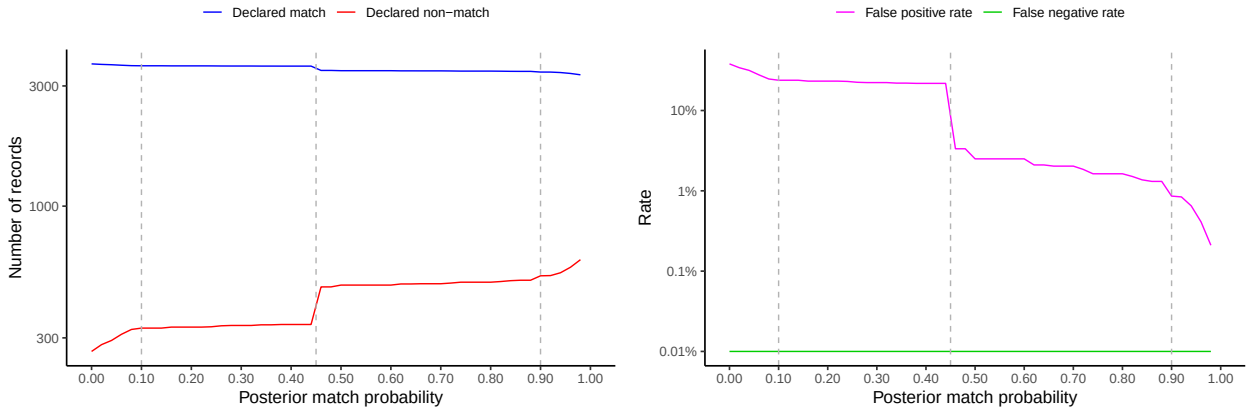
All other parameters were left as the default (see fastLink documentation).

0.2.2 Determine thresholds

We plot the posterior probabilities (zeta) and their corresponding Fellegi-Sunter weights, as demonstrated by Weber. *Note that the y-axes values are displayed as the natural logarithm of 1 plus the number of records at each zeta or weight.*



We then plot the frequencies of matches and non-matches, and the false positive rate (aka false detection rate [FDR]) and false negative rate (FNR) across the range of probability thresholds.

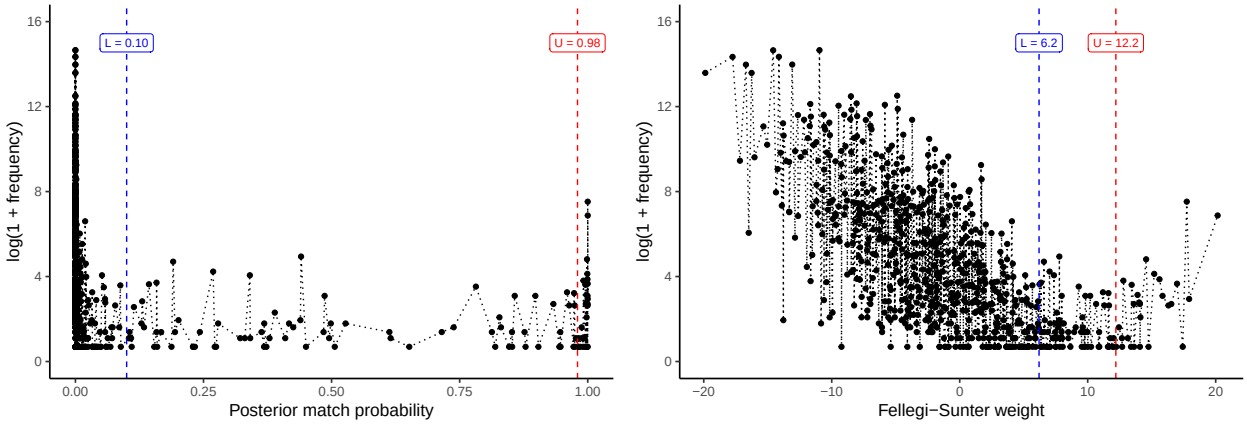


Considering the graphs, there appears to be an abrupt change in the number of matches vs. non-matches, and the FDR at zetas of ~0.10, ~0.45 and ~0.90 (*dotted lines*). The FNR appears essentially constant at 0.01% across all values of zeta.

It is most important to minimise the FDR (i.e. minimise the likelihood of declaring records a match when they are not a true match). Therefore, we set the threshold for declared matches very high (**zeta = 0.98**), which yielded an FDR <0.5%. We set the lower threshold (for declaring potential matches requiring manual review) at **zeta = 0.10**, based on the abrupt changes in the above graphs at this point.

- Zeta = 0.10 corresponds to Fellegi-Sunter weight = ~6.2.
- Zeta = 0.98 corresponds to Fellegi-Sunter weight = ~12.2.

Below are the zeta and Fellegi-Sunter weight plots with thresholds superimposed:



The chosen thresholds result in the following confusion tables:

```
## $lower_thres
## $lower_thres$confusion.table
##                       'True' Matches 'True' Non-Matches
## Declared Matches            3504.46             102.54
## Declared Non-Matches           0.27             327.73
##
## $lower_thres$addition.info
##                                 results
## Max Number of Obs to be Matched 3935.00
## Sensitivity (%)                   99.99
## Specificity (%)                   76.17
## Positive Predicted Value (%)      97.16
## Negative Predicted Value (%)      99.92
## False Positive Rate (%)           23.83
## False Negative Rate (%)            0.01
## Correctly Classified (%)          97.39
## F1 Score (%)                      98.55
##
##
## $upper_thres
## $upper_thres$confusion.table
##                       'True' Matches 'True' Non-Matches
## Declared Matches            3320.70               1.30
## Declared Non-Matches           0.45             612.55
##
## $upper_thres$addition.info
##                                 results
## Max Number of Obs to be Matched 3935.00
## Sensitivity (%)                   99.99
## Specificity (%)                   99.79
## Positive Predicted Value (%)      99.96
## Negative Predicted Value (%)      99.93
## False Positive Rate (%)            0.21
## False Negative Rate (%)            0.01
## Correctly Classified (%)          99.96
## F1 Score (%)                      99.97
```

This results in **285 records for manual review.** We deemed this to be an acceptable and pragmatic

number of records to review manually.

0.2.3 Get matches and potential matches at chosen thresholds

We subset the linkage data frames to return a data frame of matches and a data frame of potential matches using the `fastLink::getMatches()` function.

```
matches_list <- vector("list")

# With zeta >0.98 (matches)
matches_list$low <- getMatches(
  dfA = adm_link,
  dfB = dis_link,
  fl.out = matches_out,
  threshold.match = 0.98,
  combine.dfs = FALSE
)

# With zeta >0.10 (matches + potential matches)
matches_list$high <- getMatches(
  dfA = adm_link,
  dfB = dis_link,
  fl.out = matches_out,
  threshold.match = 0.10,
  combine.dfs = FALSE
)

# Session IDs for potential matches
matches_list$potential_adm <-
  matches_list$high$dfA.match$session[!matches_list$high$dfA.match$session
    %in% matches_list$low$dfA.match$session]

matches_list$potential_dis <-
  matches_list$high$dfB.match$session[!matches_list$high$dfB.match$session
                                      %in% matches_list$low$dfB.match$session]
```

We build this into a full data frame with all Neotree variables for matches and potential matches by merging on session ID (which uniquely identifies each completed admission or outcome form). N.B. `adm` and `dis` are the complete data frames of Neotree admission forms and outcome forms, respectively.

```
# Build into data frames
# Designated matches from fastLink
matches_list$matches <- tibble(
  Admission.session = matches_list$low$dfA.match$session,
  Discharge.session = matches_list$low$dfB.match$session
) %>%
  merge(adm, by = "Admission.session") %>%
  merge(dis, by = "Discharge.session")

# Potential matches from fastLink
matches_list$potentials <- tibble(
  Admission.session = matches_list$potential_adm,
  Discharge.session = matches_list$potential_dis
```

```
) %>%
  merge(adm, by = "Admission.session") %>%
  merge(dis, by = "Discharge.session")
```

There are **3322 declared matches, 285 declared potential matches, and 328 non-matches** from the Fellegi-Sunter linkage algorithm.

0.3 Manual review of potential matches

Potential matches are manually reviewed to determine their true match status. We used several factors to make a clinical judgement, including:

- Admission and outcome UIDs - any discrepancies are plausible (e.g. likely to represent a typographical error).
- Admission date and outcome (discharge or death) date - congruent and plausible.
- Admission reason/diagnosis and discharge diagnosis or cause of death - congruent.
- A review of all other variables looking for any unique features on the admission and outcome form that might indicate a true match.

From manual review of the potential matches, we decided that **258** were true matches. Thus, there were **3580** declared matches at this stage.

0.3.1 Quality checks

Finally, we performed several additional 'quality checks' to identify false-positive matches or other irregularities.

First, we checked for duplicate admission or outcome session IDs (i.e. duplicate completed admission or outcome forms) in the final linked dataset.

- Duplicated admission forms: $n = 0$
- Duplicated outcome forms: $n = 0$

Therefore, the one-to-one matching constraint was successful.

Next, we checked for duplicate admission or outcome UIDs in the final linked dataset.

- Duplicated admission UIDs: $n = 86$
- Duplicated outcome UIDs: $n = 108$

Looking at an extract from these duplicates, it was clear that they are unique babies with, for example, different birth weights, gestational ages, admission reasons, discharge diagnosis or outcome despite the same UID.

Finally, we checked for invalid admission durations (i.e. cases where the outcome date came before the admission date, or where the interval was unusually long).

Acceptable discrepancies:

- Outcome date ≤ 1 day prior to admission date - this could occur if the admission form was completed retrospectively shortly after the outcome form (e.g. if a baby was deceased on or soon after arrival to the neonatal unit).

Unacceptable discrepancies:

- Outcome date $>$ 1 day prior to admission date
- Admission duration shown to be $>$ 4 months - this is not a plausible admission duration for the neonatal unit at Sally Mugabe Central Hospital.

Distribution of admission durations:

```
##                          Min.                          1st Qu.
##           "-5d -21H -45M -2S"             "1d 6H 51M 52.25S"
##                        Median                            Mean
##               "2d 13H 2M 39S" "5d 6H 57M 15.7974860329414S"
##                       3rd Qu.                            Max.
##           "5d 14H 40M 59.25S"            "309d 14H 56M 13S"
```

- Outcome date prior to admission date: $n = 49$
  - $\leq$ 1 day prior: $n = 47$
  - $>$ 1 day prior: $n = 2$
- Admission duration shown to be $>$4 months: $n = 1$

We changed the status of these 3 cases to "non-match", to err on the side of caution. We felt these most likely represented false-positive matches.

```
## [1] "session 70565" "session 21083"
```

```
## [1] "session 10707"
```

New distribution of admission durations:

```
##                          Min.                        1st Qu.
##             "-16H -59M -25S"               "1d 6H 54M 10S"
##                       Median                          Mean
##              "2d 13H 7M 59S" "5d 5H 1M 47.4453452615999S"
##                       3rd Qu.                          Max.
##             "5d 14H 40M 46S"            "84d 22H 15M 49S"
```

A total of 3577 record pairs were thus included in the final linked dataset.

0.3.2 Flow diagram

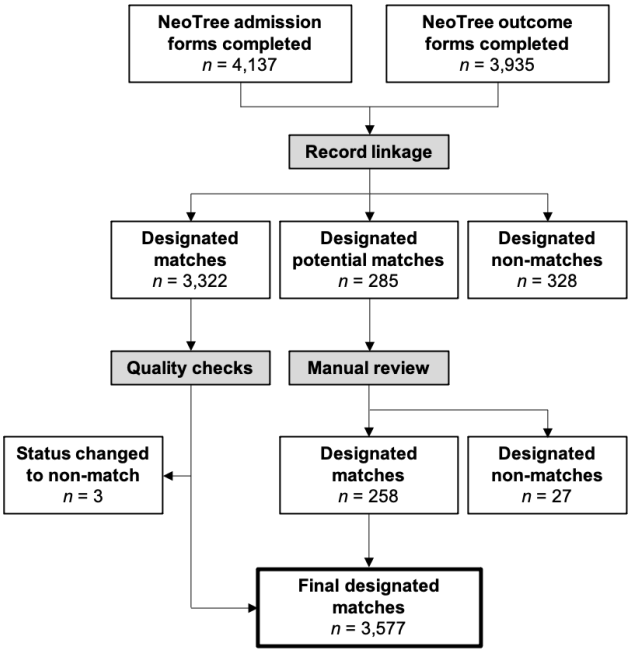


Figure 1: Flow diagram summarising record linkage