# 3-PreliminaryCleaning

We applied several preliminary cleaning steps to the raw imported data.

- Number of rows in raw admission data frame = 99468
- Number of rows in raw outcome data frame = 105139

## 0.1 Removing duplicate entries

We defined exact duplicates as entries where values for all variables were identical to one or more other entries. This occurs when data are exported from a study tablet before previous data have been erased, resulting in some entries being exported in duplicate.

Number of duplicate entries:

```
## # A tibble: 2 x 2
##   form      duplicates
##   <chr>          <int>
## 1 admission      94801
## 2 outcome       100476
```

## 0.2 Recoding missing values

We recoded empty cells or cells containing strings that signify missingness as missing values using the following custom function:

```
## function (x)
## {
##     strings <- c("", "na", "n/a", "N/A", "NA", "Nil", "nil",
##         "-")
##     x[x %in% strings] <- NA
##     x
## }
```

## 0.3 Standardising variables between admission & outcome forms

We standardised `mode of delivery` and `sex` between admission and outcome forms, so they can be used for record linkage.

Labels before standardisation:

```
## $`Mode of delivery (admission)`
## [1] "1" "2" "3" "4" "5" "6"
##
## $`Mode of delivery (outcome)`
```

```
## [1] "ECS"  "ElCS" "For"  "SVD"  "Vent"
##
## $`Sex (admission)`
## [1] "F"  "M"  "NS"
##
## $`Sex (outcome)`
## [1] "F" "M" "U"
```

Labels after standardisation:

```
## $`Mode of delivery (admission)`
## [1] "ECS"  "ElCS" "For"  "SVD"  "Vent"
##
## $`Mode of delivery (outcome)`
## [1] "ECS"  "ElCS" "For"  "SVD"  "Vent"
##
## $`Sex (admission)`
## [1] "F" "M" "U"
##
## $`Sex (outcome)`
## [1] "F" "M" "U"
```

## 0.4 Removing entries without a healthcare worker identifier

We removed entries that had not been 'signed off' by a healthcare worker with their healthcare worker identifier (HCW ID) (commonly their initials). Entries without a HCW ID occur for several reasons, e.g. (1) a healthcare worker accidentally exits the app and starts a new form upon reopening it; (2) a healthcare worker is demonstrating how to use the app to another user so does not want to mark the form as a genuine entry.

Number of entries without a HCW ID:

```
## # A tibble: 2 x 2
##   form      `no HCW ID`
##   <chr>           <int>
## 1 admission         100
## 2 outcome            88
```

## 0.5 Removing outcome form entries with invalid unique identifiers

Invalid UIDs were:

```
## # A tibble: 4 x 2
##   format                                freq
##   <chr>                                <int>
## 1 missing values                          24
## 2 strings of only zeros                   84
## 3 strings shorter than 4 characters long  12
## 4 strings containing words                 3
```

## 0.6 Limiting entries to the study period

We removed entries outwith the study period. This included entries prior to 01/02/2019, which constituted the 'pilot period' of data collection for the Neotree at SMCH.

Data import and preliminary cleaning resulted in one data frame for admission forms and one data frame for outcome forms.

- Number of rows in final admission data frame = 4137
- Number of rows in final outcome data frame = 3935

## 0.7 Flow diagram

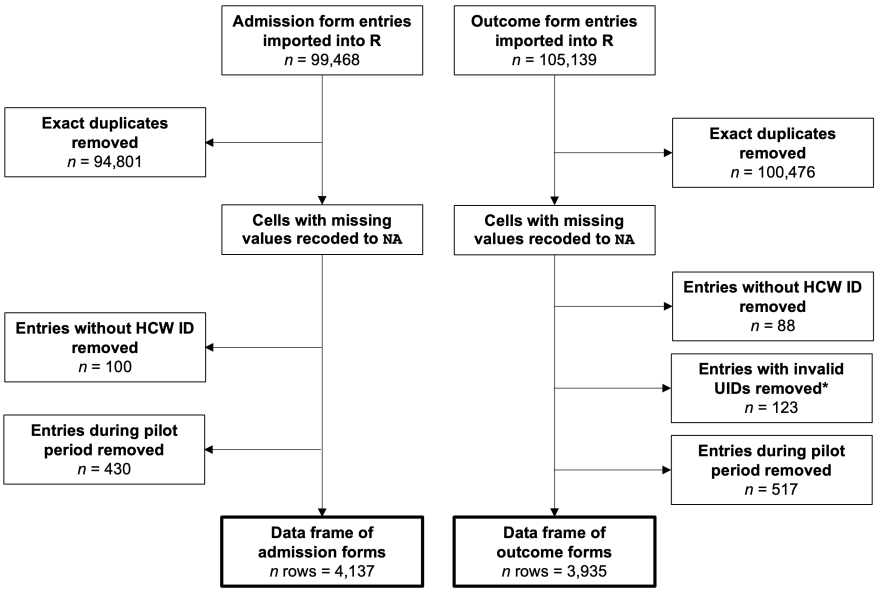


Figure 1: Flow diagram summarising preliminary data cleaning