5-FurtherCleaning

Elaboration on creating/extracting relevant variables required for model development from the Neotree dataset at Sally Mugabe Central Hospital.

0.1 Data collected by admission forms

There are 7 sections to the Neotree admission form at SMCH, Zimbabwe.

1. Emergency triage & vital signs
2. Patient information
3. Examination
4. Symptom review
5. Place of origin
6. Maternal history
7. Provisional diagnoses

N.B. Other data are collected by the app, but only relevant variables are detailed here.

0.1.1 Emergency triage & vital signs

The variables to be subset/created from this section are as follows:

Parent variable	New variable(s)	Comments
Admission.DangerSigns	et_grunt	"Grun" (yes/no)
" "	et_cyanosis	"Cyan" (yes/no)
" "	et_seizures	"Conv" (yes/no)
Admission.RR	et_rr	(numeric)
Admission.HR	et_hr	(numeric)
Admission.Temperature	et_temp	(numeric)
Admission.BW	et_bw	(numeric)
Admission.AW	*informs* et_bw	(numeric)

0.1.1.1 Admission.DangerSigns

Categorical variable with four levels:

- Grun = "Grunting or severe chest indrawings"
- Cyan = "Central cyanosis"
- Conv = "Convulsions or twitchings"
- None

Recoded into three separate variables: `et_grunt`, `et_cyanosis` and `et_seizures`.

```
## [1] "Original variable"
```

```
##
##           Conv      Conv,Grun           Cyan           Grun Grun,Conv,Cyan
##             11              3             58           1067               1
##      Grun,Cyan           None           <NA>
##             82           2353              2
```

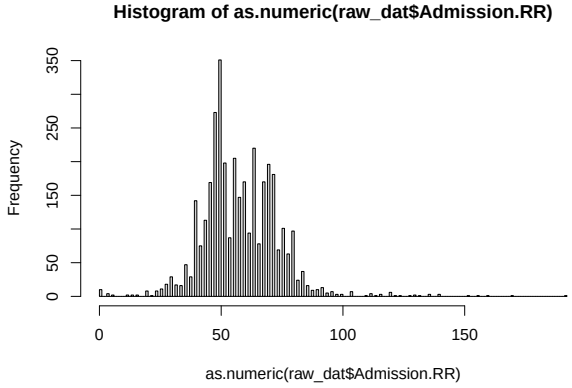
```
## [1] "New variables"
```

```
##  et_grunt    et_cyanosis et_seizures
##  no  :2422   no  :3434   no  :3560
##  yes :1153   yes : 141   yes :  15
##  NA's:   2   NA's:   2   NA's:   2
```

0.1.1.2 Admission.RR Continuous variable measured in breaths per minute.

- Some recorded values were very low (i.e. <20 breaths per minute).
  - On inspection, most died suggesting the recorded values were correct.
  - Some neonates were recorded as surviving to discharge with an initial RR < 10, despite receiving no resuscitation. This is implausible and their RR was set to missing.
- Similarly, some recorded values were very high (i.e. >100 breaths per minute).
  - After reviewing the distribution, we truncated these data to the 99.5th percentile, setting greater values to missing.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   48.00   56.00   58.32   68.00  192.00       6
```

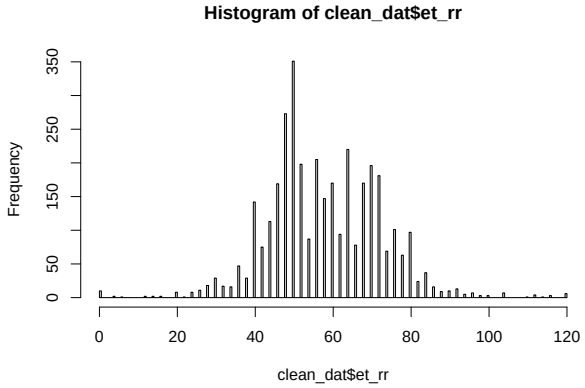



```
##    0%  0.5%    1%   50%   99% 99.5%  100%
##   0.0  13.7  24.0  56.0 104.0 120.0 192.0
```

```
## [1] "Lowest 10 values"
```

```
##
##  0  4  6 12 14 16 20 22 24 26
## 10  4  2  2  2  2  8  1  8 11
```

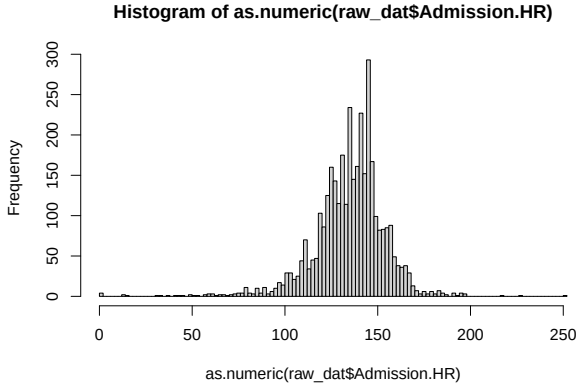
```
## [1] "Highest 20 values"

##
##  98 100 104 110 112 114 116 120 122 124 128 130 132 136 140 152 156 160 170 192
##   3   3   7   1   4   1   3   6   1   1   1   2   1   3   3   1   1   1   1   1

## [1] "Cases where RR <20"

## # A tibble: 22 x 3
##    Admission.RR Admission.Resus     Discharge.NeoTreeOutcome
##           <dbl> <chr>               <chr>
##  1           16 Stim,O2             NND
##  2            4 Stim,BVM,O2,Suc     NND
##  3            0 Stim,O2,Suc         NND
##  4            0 Stim,CPR,O2,BVM,Suc NND
##  5            6 Stim,CPR,O2,BVM,Suc NND
##  6           12 Stim,O2,BVM         NND
##  7            0 Stim,CPR,BVM        NND
##  8            6 None                DC
##  9           16 Stim,O2             NND
## 10            0 CPR,Suc,O2,BVM      NND
## 11            0 Stim,CPR,O2,BVM,Suc NND
## 12           12 Stim,BVM,O2,Suc     NND
## 13            0 Stim,BVM,O2,Suc     NND
## 14            4 None                DC
## 15           14 Stim,CPR,O2,BVM,Suc NND
## 16            4 Stim,CPR,O2,BVM,Suc NND
## 17            0 None                NND
## 18            0 Stim,CPR,O2,BVM,Suc NND
## 19            0 Stim,CPR,BVM        NND
## 20            4 None                DC
## 21           14 Stim,O2             DC
## 22            0 None                NND

## [1] "New variable"

##      et_rr
##  Min.   :  0.00
##  1st Qu.: 48.00
##  Median : 56.00
##  Mean   : 57.96
##  3rd Qu.: 68.00
##  Max.   :120.00
##  NA's   :26
```



0.1.1.3 Admission.HR Continuous variable measured in beats per minute.

- Some recorded values were very low (i.e. <50 beats per minute).
  - On inspection, most died suggesting the recorded values were correct.
  - Some neonates were recorded as surviving to discharge with an initial HR < 20, despite receiving no resuscitation. This is implausible and their HR was set to missing.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   125.0   138.0   135.6   146.0   252.0
```




```
##      0%    0.1%      1%     50%     99%   99.9%    100%
##   0.000   8.064  74.760 138.000 179.000 198.000 252.000

## [1] "Lowest 10 values"

##
##  0 14 15 32 34 38 42 43 45 50
##  4  2  1  1  1  1  1  1  1  2

## [1] "Highest 20 values"
```

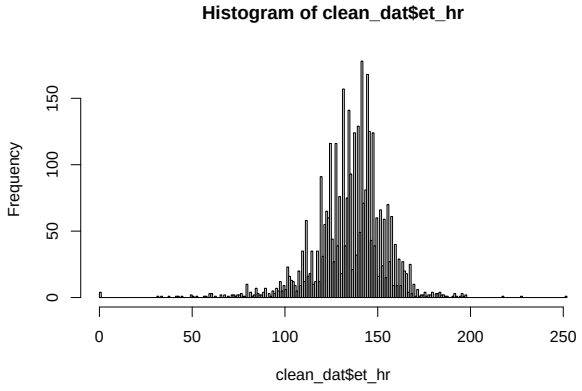
```
##
## 178 179 180 182 183 184 185 186 187 188 191 192 193 195 196 197 198 218 228 252
##   2   2   4   3   3   4   2   2   1   1   1   3   1   1   3   1   2   1   1   1
```

```
## [1] "Cases where HR <50"
```

```
## # A tibble: 13 x 4
##    Admission.HR Admission.RR Admission.Resus     Discharge.NeoTreeOutcome
##           <dbl> <chr>        <chr>               <chr>
##  1           32 60           O2,Suc              NND
##  2           14 70           None                DC
##  3            0 0            Stim,CPR,BVM        NND
##  4           38 26           Stim,CPR,O2,BVM     NND
##  5           42 44           Stim,CPR,O2,BVM,Suc NND
##  6           14 50           None                DC
##  7            0 20           Stim,CPR,O2,BVM,Suc NND
##  8            0 0            Stim,BVM,O2,Suc     NND
##  9           15 48           None                DC
## 10           43 4            Stim,CPR,O2,BVM,Suc NND
## 11           34 0            None                NND
## 12           45 26           Stim,CPR,BVM,Suc    NND
## 13            0 0            None                NND
```

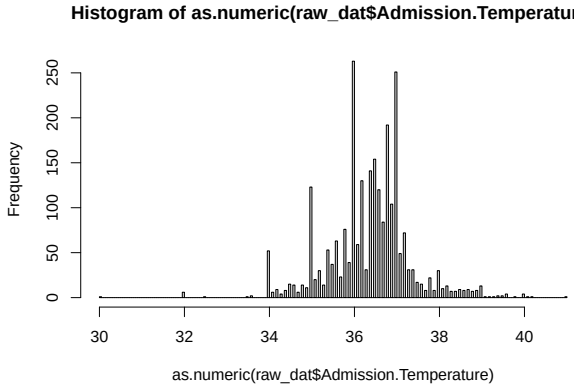
```
## [1] "New variable"
```

```
##      et_hr
##  Min.   :  0.0
##  1st Qu.:125.0
##  Median :138.0
##  Mean   :135.7
##  3rd Qu.:146.0
##  Max.   :252.0
##  NA's   :3
```



0.1.1.4 Admission.Temperature Continuous variable measured in degrees Celsius (to 0.1 precision).

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   30.00   36.00   36.50   36.38   37.00   41.00    1027
```




```
##     et_temp
##  Min.   :30.00
##  1st Qu.:36.00
##  Median :36.50
##  Mean   :36.38
##  3rd Qu.:37.00
##  Max.   :41.00
##  NA's   :1027
```

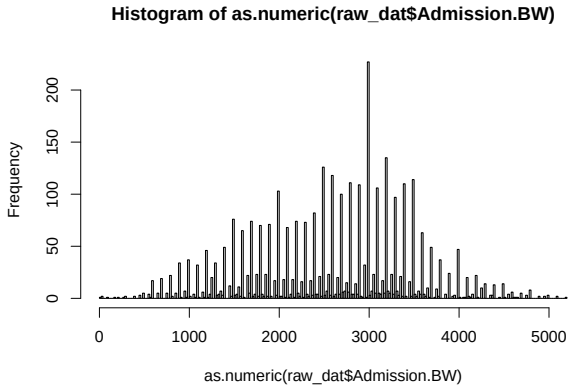
0.1.1.5 Admission.BW & Admission.AW

Continuous variables measured in grams.

- Looking at the distributions of birth weight (BW) and admission weight (AW), some values are clearly invalid.
- It is important not to assume what these values should be (e.g., for "100" the true value may have been "1000", or perhaps "3100").

```
## [1] "Distribution of birth weight"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       2    1950    2700    2592    3200    5200      48
```

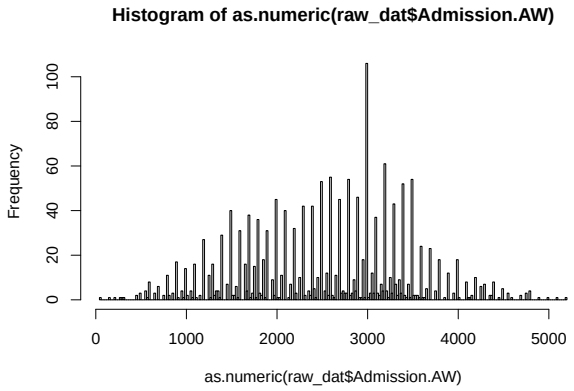



```
## [1] "Lowest values"
```

```
##
##   2  35  36 100 180 220 270 300 400 450 500 550 580 600 650 690 700 750 800 805
##   1   1   1   1   1   1   1   2   2   3   5   4   1  17   5   1  18   5  22   2
## 850 900 920 945 950 955
##   5  34   1   1   5   1
```

```
## [1] "Distribution of admission weight"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      40    1850    2600    2544    3200    5200    1877
```




```
## [1] "Lowest values"
```

```
##
##  40 150 220 280 300 310 450 500 550 580 600 650 690 700 750 800 805 850 900 920
##   1   1   1   1   1   1   2   3   4   1   8   3   1   5   2  11   2   3  17   1
## 950
##   3
```

```
## [1] "Cases where BW or AW <500g"
```

```
## # A tibble: 18 x 2
##    Admission.BW Admission.AW
##           <dbl>        <dbl>
##  1         1500          150
##  2          100           NA
##  3          300          300
##  4         3100          310
##  5          300         3000
##  6          450          450
##  7          450           NA
##  8          400           NA
##  9         2800          280
## 10          450          450
## 11          700           40
## 12           35         3375
## 13            2           NA
```

```
## 14          400           NA
## 15           36         1340
## 16          270           NA
## 17          180         1800
## 18          220          220
```

Therefore, we assessed how many cases have BW and/or AW missing, and whether it is necessary to have two weight variables (i.e., do BW and AW substantially differ?):

```
## [1] "Birth weight missing"
```

```
## [1] 48
```

```
## [1] "Admission weight missing"
```

```
## [1] 1877
```

```
## [1] "Birth weight missing but admission weight NOT missing"
```

```
## [1] 28
```

```
## [1] "Cases where BW and AW differ (and AW not missing)"
```

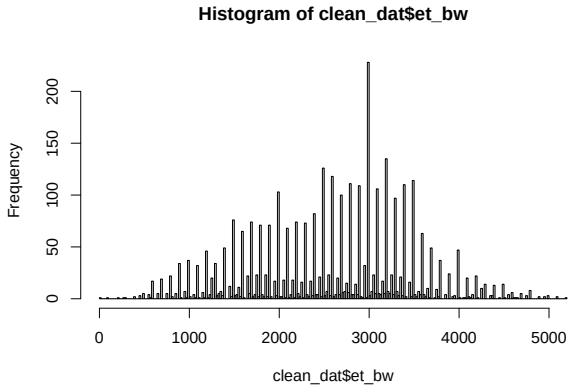
```
## # A tibble: 32 x 2
##    Admission.BW Admission.AW
##           <dbl>        <dbl>
##  1         1500          150
##  2         3780         3300
##  3         3100          310
##  4         2120         2100
##  5         1700         1660
##  6         1650         1700
##  7         1540         1890
##  8         2488         2408
##  9          300         3000
## 10         3400         3408
## 11         1500         1350
## 12         2630         2603
## 13         3300         3700
## 14         1700         1550
## 15         1660         1600
## 16         3000         2880
## 17         2700         2600
## 18         1800         1500
## 19         1320         1750
## 20         4800         4560
## 21         1500         1600
## 22         2000         2200
## 23         4000         3900
## 24         2800          280
## 25         1900         1980
## 26         2910         2700
```

```
## 27          3100          3400
## 28           700            40
## 29            35          3375
## 30          1470          1275
## 31            36          1340
## 32           180          1800
```

- There are only 32 cases where BW and AW differ.
  - Examining these cases, the differences are relatively small (excluding cases where the value is obviously erroneous). Therefore, it is unnecessary to have a separate variable for AW, and BW will suffice.
- Some values were recorded as very low (i.e. <500g).
  - If BW is consistent with the gestational age, the original value is retained.
  - If BW is inconsistent with gestational age but AW is consistent, then `et_bw` takes the value of AW.
  - Otherwise, if neither BW or AW consistent with gestational age (or AW missing), then original BW value retained and case will be excluded based on inclusion/exclusion criteria for birth weight (see below).
  - *N.B. We used the UK-WHO Neonatal and Infant Close Monitoring Growth Chart 2009 to determine weights consistent with each gestational age.*

```
## [1] "New variable"
```

```
##      et_bw
##  Min.   :   2
##  1st Qu.:1950
##  Median :2700
##  Mean   :2595
##  3rd Qu.:3200
##  Max.   :5200
##  NA's   :48
```



0.1.2 Patient information

The variables to be subset/created from this section are as follows:

Parent variable	New variable(s)	Comments
Admission.AdmReason	*informs* pi_bba	"BBA" (yes/no)
" "	pi_admreason	*takes original values* (factor)
Admission.UID	adm_uid	(string)
Admission.session	adm_session	(string)
Admission.DateTimeAdmission	adm_datetime	(date-time)
Admission.Gender	pi_sex	*takes original values* (factor)
Admission.AgeA/B/Cat/C	pi_age	(numeric)
Admission.TypeBirth	pi_type	(factor)
Admission.Gestation	pi_gest	(numeric)

0.1.2.1 Admission.AdmReason Categorical variable with many levels. No changes made to original data.

```
##
##      AD     Apg      BA     BBA    Cong    Conv     DIB      DU      FD     Fev
##      25     392     141     139      46       7     511      10      63     132
##       G    HIVX       J     LBW     Mac     Mec     NTD       O      OM    Prem
##      96      10     155     246     128     240      25     261      18     149
## PremRDS    Risk    Safe     SPn    <NA>
##     443      86     251       3       0
```

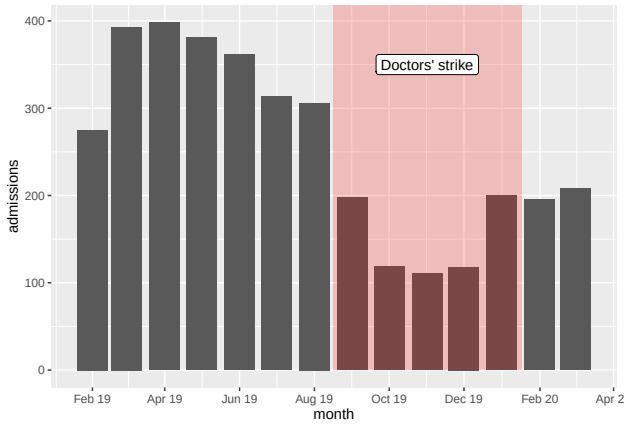
```
##   pi_admreason
##  DIB    : 511
##  PremRDS: 443
##  Apg    : 392
##  O      : 261
##  Safe   : 251
##  LBW    : 246
##  (Other):1473
```

0.1.2.2 Admission.UID & Admission.session String variables. No changes made to original data.

- Admission.UID = the unique identifier for each baby, automatically generated by the Neotree app when a new admission form is created.
- Admission.UID_alphanum = Admission.UID but with non-alphanumeric characters removed. Used for record linkage.
- Admission.session = a unique number assigned to each row of data when imported from the raw JSON files (i.e., `seq_along(1:nrow(data))`). Can be used to merge columns from the other data frames if needed in later analyses.

0.1.2.3 Admission.DateTimeAdmission String variable representing a date. Converted to POSIXct object.

- The period prior to 1st February 2019 was a 'pilot period'.
  - During this period, healthcare workers were becoming accustomed to the Neotree app and only a subset of admissions and outcomes were recorded.



0.1.2.4 Admission.Gender

Categorical variable with three levels.

- Male
- Female
- Unsure

No changes made to original data.

```
##
##    F    M    U <NA>
## 1608 1965    4    0
```

```
## [1] "New variable"
```

```
##  pi_sex
##  f:1608
##  m:1965
##  u:   4
```

0.1.2.5 Admission.AgeA/B/Cat/C

Categorical or string variables representing age at admission:

- Admission.AgeA = Is the baby aged less than 1 week?
  - Binary categorical variable: yes (Y) or no (N)
- Admission.AgeB = If AgeA = yes, the baby's age to the nearest hour
  - String varible in the format `X days, Y hours`
- Admission.AgeCat = If AgeA = yes, the age category that the baby falls into
  - Categorical variable with 5 levels:
    - Fresh newborn (<2 hours-old)
    - Newborn 2-23 hours-old
    - Newborn 24-47 hours-old
    - Infant 48-71 hours-old
    - Infant 72 hours-old
- Admission.AgeC = If AgeA = no, the baby's age to the nearest day

– String variable in the format `X days`

N.B. If the reason for admission is "dumped baby", then age is not recorded.

```
## # A tibble: 4 x 3
##   variable         n_miss pct_miss
##   <chr>             <int>    <dbl>
## 1 Admission.AgeC     3557   99.4
## 2 Admission.AgeB      657   18.4
## 3 Admission.AgeCat    109    3.05
## 4 Admission.AgeA       13    0.363
```

```
## [1] "Missing both AgeB and AgeCat"
```

```
## [1] 28
```

All age variables have a high proportion of missingness except Admission.AgeCat and Admission.AgeA.

- Since Admission.AgeA is a simple binary question of whether the baby is less than one week-old, using Admission.AgeCat is more informative.
- This means age will be a categorical variable rather than a continuous variable, but this is preferable to reduce the proportion of missing values.

We can transform Admission.AgeB into a continuous variable of age in hours, and then check to ensure Admission.AgeB is congruent with Admission.AgeCat:

```
## [1] "Admission.AgeB, original"
```

```
##  [1] "18 hours"         "1 day, 9 hours"   "1 hour"           "1 day, 9 hours"
##  [5] "16 hours"         "1 day, 5 hours"   "1 day, 5 hours"   "1 day, 15 hours"
##  [9] "6 hours"          "15 hours"         "13 hours"         "2 hours"
## [13] "19 hours"         "4 hours"          "11 hours"         "2 days, 18 hours"
## [17] "14 hours"         "11 hours"         "1 day, 3 hours"   NA
```

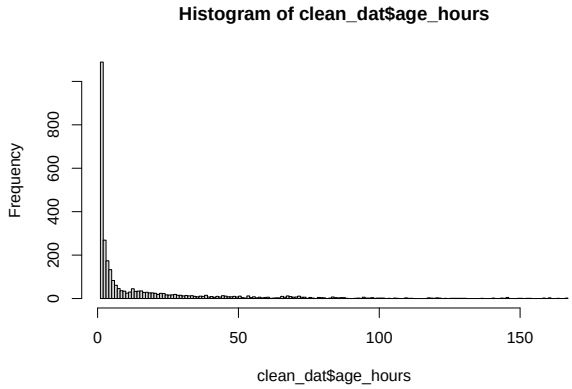
```
## [1] "Note some anomalies: negative values or >1 week-old"
```

```
##  [1] "-21 hours"               "-10 hours"
##  [3] "1 month5 days, 16 hours" "1 month1 day, 4 hours"
##  [5] "-17 hours"               "-10 hours"
##  [7] "-3 hours"                "-18 hours"
##  [9] "-5 hours"                "-11 hours"
## [11] "-23 hours"               "-20 hours"
## [13] "-20 hours"               "-23 hours"
## [15] "-23 hours"               "-23 hours"
## [17] "-20 hours"               "-5 hours"
## [19] "-10 hours"               "1 month4 days, 7 hours"
## [21] "-6 hours"                "-22 hours"
## [23] "-21 hours"               "-22 hours"
## [25] "-5 hours"                "-9 hours"
## [27] "-8 hours"                "-10 hours"
```

```
## [1] "Check this new variable, age in hours"
```

```
## # A tibble: 10 x 2
##    Admission.AgeB  age_hours
##    <chr>                <dbl>
##  1 18 hours                18
##  2 1 day, 9 hours          33
##  3 1 hour                   1
##  4 1 day, 9 hours          33
##  5 16 hours                16
##  6 1 day, 5 hours          29
##  7 1 day, 5 hours          29
##  8 1 day, 15 hours         39
##  9 6 hours                  6
## 10 15 hours                15
```

```
##    age_hours
##  Min.   :  1.00
##  1st Qu.:  2.00
##  Median :  4.00
##  Mean   : 15.61
##  3rd Qu.: 18.00
##  Max.   :167.00
##  NA's   :685
```




```
## [1] "Generate agecat_new based on age_hours values"
```

```
## [1] "Cases where agecat != agecat_new"
```

```
## [1] 259
```

```
## # A tibble: 6 x 5
##   Admission.AgeA age_hours agecat agecat_new Admission.AgeC
##   <chr>              <dbl> <chr>  <fct>      <chr>
## 1 Y                     33 NB24   NB48       <NA>
## 2 Y                     27 NB24   NB48       <NA>
## 3 Y                     95 INF72  INF        <NA>
## 4 Y                      1 NB24   FNB        <NA>
## 5 Y                     67 INF    INF72      <NA>
## 6 Y                     38 NB24   NB48       <NA>
```

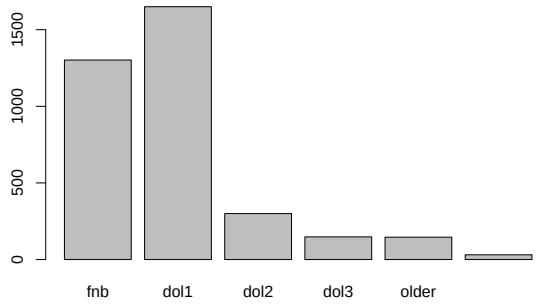
There are some discrepancies between the age from Admission.AgeB (automatically generated by the app from date-time of birth and admission date-time) and the age category selected by the healthcare workers (recorded as Admission.AgeCat).

- These discrepancies occur in relatively few cases and likely represent a misunderstanding of the age category definitions by healthcare workers using the app.
- As Admission.AgeB is generated automatically by the app, it is less liable to errors than Admission.AgeCat.
- Therefore, the following rules will be applied to create the age variable:
  - Where Admission.AgeB is *not* missing, we use this variable to assign the age category.
  - Where Admission.AgeB is missing but Admission.AgeCat is *not* missing, we use the value of Admission.AgeCat.
  - Where Admission.AgeCat is missing but Admission.Age == "N", then the baby is older than one week, so is assigned to the "infant" category.
  - Where all the above are missing, the new age variable is missing.

```
## [1] 14
```

```
## # A tibble: 14 x 4
##    Admission.AgeA Admission.AgeB  agecat pi_age
##    <chr>          <chr>           <chr>  <chr>
##  1 N              15 hours        <NA>   NB24
##  2 N              2 days, 6 hours <NA>   INF72
##  3 N              2 days, 3 hours <NA>   INF72
##  4 N              14 hours        <NA>   NB24
##  5 N              1 day, 17 hours <NA>   NB48
##  6 N              21 hours        <NA>   NB24
##  7 N              1 hour          <NA>   FNB
##  8 N              1 day, 3 hours  <NA>   NB48
##  9 N              1 day, 20 hours <NA>   NB48
## 10 N              16 hours        <NA>   NB24
## 11 N              1 day, 3 hours  <NA>   NB48
## 12 N              1 day           <NA>   NB48
## 13 N              1 day, 2 hours  <NA>   NB48
## 14 N              1 day, 11 hours <NA>   NB48
```

```
##    pi_age
##  fnb  :1302
##  dol1 :1650
##  dol2 : 300
##  dol3 : 148
##  older: 146
##  NA's :  31
```



There are several cases where Admission.AgeA would suggest the baby is ≥ 1 week-old, yet Admission.AgeB (and, thus, pi_age) does not correlate with this. Admission.AgeB is likely the most accurate source of age and so this value will be used.

0.1.2.6 Admission.TypeBirth Categorical variable with six levels:

- Singleton
- Twin number 1
- Twin number 2
- Triplet number 1
- Triplet number 2
- Triplet number 3

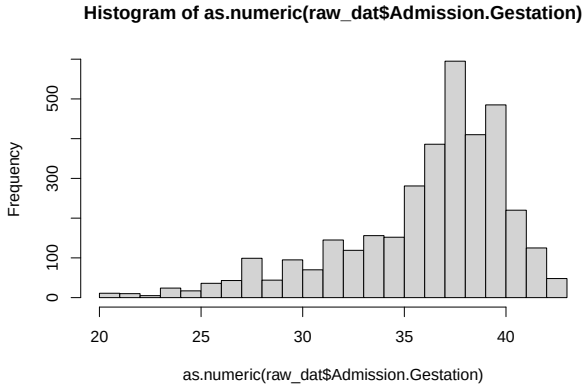
No changes made to original data.

```
##
##    S  Tr1  Tr2  Tr3  Tw1  Tw2 <NA>
## 3217    6    5    6  187  153    3
```

```
##       pi_type
##  singleton:3217
##  twin1    : 187
##  twin2    : 153
##  triplet1 :   6
##  triplet2 :   5
##  triplet3 :   6
##  NA's     :   3
```

0.1.2.7 Admission.Gestation Continuous variable measured in weeks. No changes made to original data.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   20.00   35.00   38.00   36.54   39.00   43.00       1
```




```
##     pi_gest
##  Min.   :20.00
##  1st Qu.:35.00
##  Median :38.00
##  Mean   :36.54
##  3rd Qu.:39.00
##  Max.   :43.00
##  NA's   :1
```

0.1.3 Examination

The variables to be subset/created from this section are as follows:

Parent variable	New variable(s)	Comments
Admission.Fontanelle	oe_fontanelle	*takes values* (factor)
Admission.Activity	oe_activity	*takes values* (factor)
Admission.SignsRD	oe_nasalflare	"NFL" (yes/no)
" "	oe_retractions	"CHI" (yes/no)
" "	oe_grunt	"GR" (yes/no)
Admission.WOB	oe_wob	*takes values* (factor), add "normal" if SignsRD == "None", NA if SignsRD missing
Admission.Colour	oe_colour	*takes values* (factor)
Admission.Abdomen	oe_abdodist	"Dist" (yes/no)
Admission.Umbilicus	oe_omphalitis	"Inf" (yes/no)
Admission.Skin	oe_abskin	not "None" (yes/no)

0.1.3.1 Admission.Fontanelle Categorical variable with three levels:

- Bulging = "Bulging"
- Flat = "Flat"
- Sunken = "Sunken"

No changes made to original data.

```
##
## Bulg Flat Sunk <NA>
##   16 3546   15    0
```

```
##  oe_fontanelle
##  flat   :3546
##  sunken :  15
##  bulging:  16
```

0.1.3.2 Admission.Activity

Categorical variable with five levels:

- Alert = "Alert, active, appropriate"
- Coma = "Coma (unresponsive)"
- Convulsions = "Seizures, convulsions, or twitchings"
- Irritable = "Irritable"
- Lethargic = "Lethargic, quiet, decreased activity"

No changes made to original data.

```
##
## Alert  Coma  Conv Irrit  Leth  <NA>
##  2791    48    16    77   645     0
```

```
##     oe_activity
##  alert    :2791
##  lethargic: 645
##  irritable:  77
##  seizures :  16
##  coma     :  48
```

0.1.3.3 Admission.SignsRD

Categorical variable with five levels:

- Chest retractions = "Chest in-drawings"
- Grunting = "Grunting"
- Nasal flaring = "Nasal flaring"
- ~~Gasping = "Gasping"~~
- ~~Stridor = "Stridor"~~
- ~~Head nodding = "Head nodding"~~
- ~~Tracheal tug = "Tracheal tug"~~
- None

Of these, only the first three categories are candidate predictors for this study. No changes made to original data.

```
##
##                  CHI               CHI,GR           CHI,HN,NFL
##                  268                   71                    1
##               CHI,NFL           CHI,NFL,GR                 Gasp
##                  488                  307                   45
##             Gasp,CHI          Gasp,CHI,GR         Gasp,CHI,NFL
##                   16                    7                   19
```

```
##       Gasp,CHI,NFL,GR               Gasp,GR       Gasp,HN,CHI,NFL
##                    20                     4                     1
##    Gasp,HN,CHI,NFL,GR              Gasp,NFL          Gasp,NFL,CHI
##                     3                     3                     6
##       Gasp,NFL,CHI,GR                    GR                HN,CHI
##                     4                    35                     4
##             HN,CHI,GR             HN,CHI,NFL         HN,CHI,NFL,GR
##                     5                    21                    48
##                HN,NFL             HN,NFL,CHI         HN,NFL,CHI,GR
##                     1                     1                     7
##             HN,NFL,GR                   NFL                NFL,CHI
##                     1                   189                    87
##            NFL,CHI,GR                NFL,GR                 NFL,HN
##                    31                    44                     1
##            NFL,HN,CHI              NFL,HN,GR                  None
##                     1                     1                  1788
##                    ST                ST,CHI         ST,CHI,NFL,GR
##                     1                     1                     1
##                 ST,HN                ST,NFL                    TT
##                     1                     1                     2
##                TT,CHI             TT,CHI,NFL         TT,CHI,NFL,GR
##                     4                     8                    10
##       TT,Gasp,CHI,NFL    TT,Gasp,CHI,NFL,GR TT,Gasp,HN,CHI,NFL,GR
##                     1                     2                     1
##    TT,Gasp,NFL,CHI,GR              TT,HN,CHI      TT,HN,CHI,NFL,GR
##                     1                     1                     6
##      TT,HN,NFL,CHI,GR             TT,NFL,CHI         TT,NFL,CHI,GR
##                     1                     3                     1
##         TT,NFL,HN,CHI      TT,ST,CHI,NFL,GR                  <NA>
##                     1                     1                     1

##  oe_nasalflare oe_retractions oe_grunt
##  no  :2253     no  :2117      no  :2964
##  yes :1323     yes :1459      yes : 612
##  NA's:   1     NA's:   1      NA's:   1
```

0.1.3.4 Admission.WOB

Categorical variable with three levels:

- Mildly increased work of breathing (WOB) = "Mild"
- Moderately increased WOB = "Moderate"
- Severely increased WOB = "Severe"

N.B. At the time of study, this variable was only completed if Admission.SignsRD was recorded as "nasal flaring", "chest retractions", "head nodding", "grunting", or "tracheal tug". A value was *not* entered if Admission.SignsRD was recorded as "gasping" or "stridor".

The following rules were applied to create the new WOB variable:

- NA if Admission.SignsRD is NA;
- "normal" if Admission.SignsRD == "none";
- NA if Admission.SignsRD == "gasping" or "stridor".

```
##
## Mild  Mod  Sev <NA>
##  520  885  339 1833
```

```
##       oe_wob
##  normal  :1788
##  mild    : 519
##  moderate: 885
##  severe  : 338
##  NA's    :  47
```

```
## [1] normal normal severe mild   mild   severe
## Levels: normal mild moderate severe
```

0.1.3.5 Admission.Colour

Categorical variable with four levels:

- Pink = "Pink"
- Blue = "Blue"
- White = "White"
- Yellow = "Yellow"

No changes made to original data.

```
##
##  Blue  Pink White  Yell  <NA>
##   129  3353    21    74     0
```

```
##   oe_colour
##  pink  :3353
##  pale  :  21
##  blue  : 129
##  yellow:  74
```

0.1.3.6 Admission.Abdomen

Categorical variable with eight levels:

- Distended = "Distended"
- ~~Hepatomegaly = "Hepatomegaly"~~
- ~~Splenomegaly = "Splenomegaly"~~
- ~~Abdominal mass = "Abdominal mass"~~
- ~~Gastroschisis = "Gastroschisis"~~
- ~~Omphalocele = "Omphalocele"~~
- ~~Prune belly = "Prune belly"~~
- Normal = "Soft and normal"

Of these, only abdominal distention is a candidate predictor for this study. No changes made to original data.

```
##
##             AbMass        AbMass,Dist       AbMass,PrunB               Dist
##                  4                  4                  1                 45
```

```
##         Dist,PrunB              GSchis              HepMeg        HepMeg,Dist
##                  1                  75                   2                  1
##               Norm                Omph          Omph,Norm         PrunB,Norm
##               3419                  15                   1                  5
##        SplMeg,Dist SplMeg,Dist,HepMeg               <NA>
##                  1                   1                   2
```

```
##  oe_abdodist
##  no  :3522
##  yes :  53
##  NA's:   2
```

0.1.3.7 Admission.Umbilicus

Categorical variable with four levels:

- Infected = "Red skin all around umbilicus"
- ~~Blood-stained = "Bleeding"~~
- ~~Meconium-stained = "Meconium stained"~~
- ~~Abnormal = "Abnormal looking"~~
- ~~Hernia = "Umbilical hernia"~~
- Normal = "Healthy and clean"

Of these, only omphalitis (i.e. "infected" umbilicus) is a candidate predictor for this study. No changes made to original data.

```
##
##     Abn   Abn,H      Bl    Bl,H       H     Inf Inf,Abn     Mec    Norm    <NA>
##      52       1       6       1       4      16       1      64    3432       0
```

```
##  oe_omphalitis
##  no :3560
##  yes:  17
```

0.1.3.8 Admission.Skin

Categorical variable with four levels:

- Pustules = "Pustules all over"
- Abscess = "Big boil/abscess"
- Rash = "Other skin rash"
- None = "Normal"

Due to distribution of categories, dichotomised into "abnormal skin" yes/no.

```
##
##      None      Rash Rash,PUST      <NA>
##      3540        36         1         0
```

```
##  oe_abskin
##  no :3540
##  yes:  37
```

0.1.4 Symptom review

The variables to be subset/created from this section are as follows:

Parent variable	New variable(s)	Comments
Admission.Vomiting	hx_vomit	*modified values* (factor)

0.1.4.1 Admission.Vomiting Categorical variable with five levels:

- Yes, vomiting = "Vomiting all feeds"
- Yes, green vomit = "Vomiting bright green"
- Yes, bloody vomit = "Vomiting with blood"
- Posseting = "Small milky possets after feeds (normal)"
- No vomiting = "NONE"

In the original variable, some cases were coded with multiple categories. The new variable was recoded to ensure mutually exclusive groups.

```
##
##        No      Poss       Yes Yes,YesGr     YesBl     YesGr      <NA>
##      3482        21        18         2         6        48         0
```

```
##     hx_vomit
##  no     :3503
##  yellow :  18
##  bilious:  50
##  bloody :   6
```

0.1.5 Maternal history (obstetric history)

The variables to be subset/created from this section are as follows:

Parent variable	New variable(s)	Comments
Admission.ROMlength	oh_prom2	"PROM" (yes/no)
Admission.RFSepsis	oh_prom	"PROM" (yes/no)
" "	oh_matfever	"MF" (yes/no)
" "	oh_offliquor	"OL" (yes/no)
Both of the above	co_prom	"yes" if oh_prom OR oh_prom2 == "yes" (yes/no)
Admission.ModeDelivery	oh_delivery	*takes values* (factor)

0.1.5.1 Admission.ROMlength Binary categorical variable:

- PROM = ">18 hours"
- NOPROM = "<18 hours"

No changes made to original data.

N.B. This is one of two PROM-related data points collected:

1. Admission.ROMlength (this variable)
2. Admission.RFSepsis (categorical variable with one catergory for PROM) - see below

```
##
## NOPROM   PROM   <NA>
##   1894    361   1322
```

```
##
##   no  yes <NA>
## 1894  361 1322
```

0.1.5.2 Admission.RFSepsis Categorical variable with seven levels:

- Prolonged rupture of membranes = "PROM more than 18 hrs"
- Maternal fever during labour = "Maternal fever in labour"
- Offensive liquor = "Offensive liquor"
- ~~Prematurity = "Prematurity <37 weeks"~~
- ~~Prolonged second stage of labour = "Prolonged second stage"~~
- ~~Born before arrival to hospital = "Born before arrival (BBA)"~~
- None

Of these, only the first three are candidate predictors for this study. Although prematurity is also a candidate predictor, this information is obtained more precisely from Admission.Gestation (*see above*).

No changes made to original data.

```
##
##             BBA          BBA,OL        BBA,Prem              MF     MF,BBA,Prem
##             127               1              27               8               2
##     MF,Pr2nd,OL   MF,Pr2nd,PROM         MF,Prem     MF,Prem,BBA         MF,PROM
##               1               2               4               1               3
##      MF,PROM,OL MF,PROM,Prem,OL            NONE              OL         OL,Prem
##               1               1            2029              89               2
##           Pr2nd        Pr2nd,OL      Pr2nd,Prem      Pr2nd,PROM   Pr2nd,PROM,OL
##              63              16               5              16               5
## Pr2nd,PROM,Prem            Prem        Prem,BBA         Prem,OL            PROM
##               1             773              69              11             167
##        PROM,BBA   PROM,BBA,Prem         PROM,OL    PROM,OL,Prem       PROM,Prem
##               1               1              39               2             100
##   PROM,Prem,BBA    PROM,Prem,OL            <NA>
##               2               7               1
```

```
##  oh_prom     oh_matfever oh_offliquor
##  no  :3228   no  :3553   no  :3401
##  yes : 348   yes :  23   yes : 175
##  NA's:   1   NA's:   1   NA's:   1
```

0.1.5.3 Creating a single variable to capture PROM As mentioned above, there are two PROM-related data points collected:

1. Admission.ROMlength - now `oh_prom2` from above

2. `Admission.RFSepsis == "PROM"` - now `oh_prom` from above

Recoded into a single variable with "yes" if either of the above variables suggest the presence of PROM.

```
## [1] "Compare coding & distribution between both PROM variables..."
```

```
##  oh_prom     oh_prom2
##  no  :3228   no  :1894
##  yes : 348   yes : 361
##  NA's:   1   NA's:1322
```

```
##
##         no  yes
##   no  1881   36
##   yes   12  325
```

```
## [1] "New combined variable..."
```

```
##   no  yes
## 3193  384
```

0.1.5.4 Admission.ModeDelivery

Categorical variable with five levels:

- Emergency caesarean section = "Emergency caesarean section"
- Elective caesarean section = "Elective caesarean section"
- Forceps = "Forceps extraction"
- Spontaneous vaginal delivery = "Spontaneous vaginal delivery"
- Ventouse = "Vacuum extraction"

No changes made to original data.

```
##
##  ECS ElCS  For  SVD Vent <NA>
##  726  186    1 2620   44    0
```

```
##       oh_delivery
##  svd        :2620
##  electiveCS : 186
##  emergencyCS: 726
##  forceps    :   1
##  ventouse   :  44
```

0.2 Data collected by outcome forms

There are two groups of outcome variables to consider:

1. Participant demographics
2. Model outcome data

0.2.1 Participant demographics

The variables to be subset/created from this section are as follows:

Parent variable	New variable(s)	Comments
Discharge.session	dis_session	(string)
Discharge.NeoTreeID	dis_uid	(string)
Discharge.NeoTreeOutcome	outcome	*takes values* (factor)
Discharge.DateTimeDischarge	outcome_datetime	(date-time)
Discharge.DateTimeDeath	outcome_datetime	(date-time)
several	adm_dur	(period)

0.2.1.1 Discharge.NeoTreeID & Discharge.session

String variables.

- Discharge.NeoTreeID = the unique identifier for each baby, automatically generated by the Neotree app when a new admission form is created. Entered manually by the healthcare worker completing the outcome form.
- Discharge.NeoTreeID_alphanum = the unique identifier but with non-alphanumeric characters removed. Used for record linkage.
- Discharge.session = a unique number assigned to each row of data when imported from the raw JSON files (i.e., `seq_along(1:nrow(data))`). Can be used to merge columns from other data frames if required in future analyses.

No changes made to original data.

```
##    dis_uid          dis_session
##  Length:3577        Length:3577
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

0.2.1.2 Discharge.NeoTreeOutcome

Categorical variable with five levels:

- Discharged = "Discharged"
- Death = "Died"
- Transferred within the hospital = "Transferred to other ward"
- Transferred to another hospital or facility = "Transferred to other hospital"
- Absconded = "Absconded"

Dichotomised into died/discharged. For this study, we considered a participant to be discharged if any outcome other than "death" was recorded.

```
##
##  ABS   DC  NND  TRH  TRO <NA>
##    3 2887  679    6    2    0
```

```
##        outcome
##  died      : 679
##  discharged:2898
```

0.2.1.3 Discharge.DateTimeDischarge & Discharge.DateTimeDeath

String variables representing dates.

```
## [1] "Ensure outcome matches date variable recorded..."
```

```
## [1] "Discharge.DateTimeDischarge missing..."
```

```
## [1] 2900
```

```
## [1] "Discharge.DateTimeDeath missing..."
```

```
## [1] 681
```

```
## [1] "Both missing..."
```

```
## [1] 4
```

There are 4 cases where both a discharge date and date of death are recorded. For these, we used the date corresponding to the recorded outcome.

0.2.1.4 Admission duration

It is useful to have a variable denoting the admission duration for each participant. Calculated from the admission and outcome dates.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.7079  1.2876  2.5472  5.2043  5.6116 85.0526
```

```
## [1] 48
```

There are 48 cases where admission duration is ≤ 0.

- These most likely represent errors when inputting the admission and/or outcome date.
- Although a tolerance of outcome date ≤ 1 day prior to admission date was allowed for record linkage, cases with negative admission durations were excluded from the main analysis because this anomaly questioned the accuracy of some other variables for that participant, e.g., chronological age (which is calculated automatically within the app from birth date-time and admission date-time).

0.3 Model outcome data

The primary outcome was early-onset sepsis, defined as sepsis with onset within the first 72 hours of life, as diagnosed by the treating consultant neonatologist.

Variable	Comments

0.3.1 Supporting variables

The variables required to create the outcome variable are as follows:

Variable	Comments
Discharge.DIAGDIS1	Primary discharge diagnosis
Discharge.DIAGDIS1OT	Free text field if primary discharge diagnosis == "other"
Discharge.OthProbs	Other problems during admission
Discharge.OthProbsOth	Free text field if other problems == "other"
Discharge.CauseDeath	Primary cause of death
Discharge.CauseDeathOther	Free text field if primary cause of death == "other"
Discharge.ContCauseDeath	Contributory cause(s) of death
Discharge.ContCauseDeathOth	Free text field of contributory cause of death == "other"

```
##
##    AN   BBA    BI    BO   CHD  DEHY  EONS    FD     G   HIE  HIVX HIVXH HIVXL
##     4    95    11     4     8     8   197    38     8   376    11    15    48
##  JAUN   LBW  LONS    MA   Mac    MD    NB   OCA    OM   OTH    PN    PR PRRDS
##   231   126    26   119   134     4    40    29    12   314     9   166   269
##    Ri  Safe   TTN  Twin  <NA>
##    72   220   294    11   678

##
##    ASP     CA   EONS Gastro    HIE   LONS    MAS    NEC    OTH     PN     PR
##     22     17     35     75    117     10      5      2     63      6     39
##  PRRDS   <NA>
##    288   2898

## [1] "Ensure all discharges have discharge diagnosis recorded..."

## [1] 0

## [1] 0

## [1] "Ensure all deaths have cause of death recorded..."

## [1] 0

## [1] 0

## [1] "New variables..."

##    diagnosis    diagnosis_other      diagnosis2   diagnosis2_other
##  HIE    : 376   Length:3577        NONE   :1449   Length:3577
##  OTH    : 314   Class :character   OTH    : 231   Class :character
##  TTN    : 294   Mode  :character   LBW    : 181   Mode  :character
##  PRRDS  : 269                      JAU    : 147
```

```
##  JAUN    : 231                     HIVX    :  95
##  (Other):1415                      (Other): 795
##  NA's    : 678                     NA's    : 679
##   cause_death    cause_death_other   cause_death2  cause_death2_other
##  PRRDS   : 288   Length:3577        NONE    : 221   Length:3577
##  HIE     : 117   Class :character   LBW     :  78   Class :character
##  Gastro  :  75   Mode  :character   OTH     :  45   Mode  :character
##  OTH     :  63                      PRRDS   :  33
##  PR      :  39                      EONS    :  25
##  (Other):  97                       (Other): 277
##  NA's    :2898                      NA's    :2898
```

0.3.2 Outcome variable (early-onset neonatal sepsis)

Binary categorical variable of early-onset sepsis yes/no.

First, we explored the free text fields for variations of "early-onset sepsis" that would need to be captured by the outcome variable:

```
# Explore free text (too long to print in full):

# clean_dat %>%
#   select(diagnosis_other) %>%
#   filter(grepl("sep|eons|early", diagnosis_other, ignore.case = T))
#
# clean_dat %>%
#   select(diagnosis2_other) %>%
#   filter(grepl("sep|eons|early", diagnosis2_other, ignore.case = T))
#
# clean_dat %>%
#   select(cause_death_other) %>%
#   filter(grepl("sep|eons|early", cause_death_other, ignore.case = T))
#
# clean_dat %>%
#   select(cause_death2_other) %>%
#   filter(grepl("sep|eons|early", cause_death2_other, ignore.case = T))
```

Relevant free text entries identified:

Variable	Relevant free text entries
Discharge.DIAGDIS1OT	*None*
Discharge.OthProbsOth	"Early Onset Neonatal Sepsis"
Discharge.CauseDeathOther	"Early onset neonatal sepsis", "earlyonset neonatal sepsis"
Discharge.ContCauseDeathOth	*None*

N.B. "Risk of sepsis", "unconfirmed sepsis" or "sepsis" were not included.

Next, we created the outcome variable.

```
# Create variable
clean_dat <- clean_dat %>%
  mutate(sepsis = factor(
```

```
    case_when(
      # 1. Discharge diagnosis of EONS:
      diagnosis == "EONS" ~ "yes",
      # 2. Other discharge problem includes EONS:
      grepl("EONS", diagnosis2) ~ "yes",
      grepl("Early Onset Neonatal Sepsis", diagnosis2_other) ~ "yes",
      # 3. Cause of death of EONS:
      cause_death == "EONS" ~ "yes",
      grepl(
        "Early onset neonatal sepsis|earlyonset neonatal sepsis",
        cause_death_other
      ) ~ "yes",
      # 4. Contributory cause of death includes EONS:
      grepl("EONS", cause_death2) ~ "yes",
      # Else, no diagnosis of EONS:
      TRUE ~ "no"
    )
  ))

# Check new variable
clean_dat %>%
  select(sepsis) %>%
  summary()
```

```
##  sepsis
##  no :3170
##  yes: 407
```

0.3.3 Inclusion and exclusion criteria

Our inclusion and exclusion criteria were:

Inclusion criteria	Exclusion criteria
Chronological age <72 hours	Not singletons or first-born multiples
Gestation 32+0 weeks at birth	Died at admission to the unit (HR or RR = 0)
Birth weight 1500 grams	Major congenital anomalies*
-	Anomalous admission duration (<0 days)

*Major congenital anomalies included congenital heart defects, open spina bifida, gastroschisis or omphalocele, and/or genetic syndromes.

The counts of participants excluded due to each criterion are:

```
## # A tibble: 7 x 2
##   criterion                           count
##   <chr>                               <int>
## 1 Admitted  72h of life                 146
## 2 Very premature                        454
## 3 Very low birth weight                 408
## 4 Dead on admission                      11
## 5 Not singleton or first-born multiple  164
## 6 Major congenital anomaly              182
## 7 Anomalous admission duration           47
```

0.3.4 Flow diagram of participant inclusion