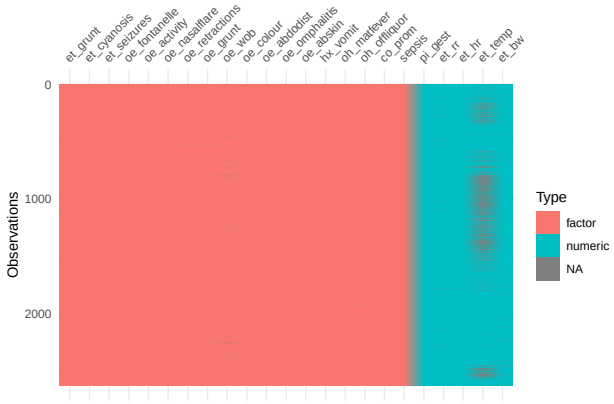
6-MissingData

Description of missing data analysis.

0.1 Assess missingness

0.1.1 Visualise data frame

A graphical representation of the data types and proportion of missing values for each variable is shown below. Ancillary variables that are not required for modelling are not shown. Variables in the data frame are plotted on the x-axis and each observation (i.e. participant) is plotted on the y-axis. Missing values are shaded grey.



0.1.2 Variable-wise missingness

The number and percentage of missing values for each variable is shown below. In total, 14 variables had missing values.

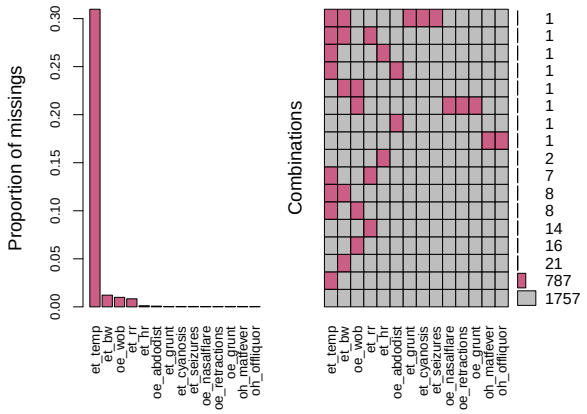
```
## # A tibble: 23 x 3
##    variable       n_miss pct_miss
##    <chr>           <int>    <dbl>
##  1 et_temp           814  31.0
##  2 et_bw              32   1.22
##  3 oe_wob             26   0.989
##  4 et_rr              22   0.837
##  5 et_hr               3   0.114
##  6 oe_abdodist         2   0.0761
##  7 et_grunt            1   0.0381
##  8 et_cyanosis         1   0.0381
##  9 et_seizures         1   0.0381
## 10 oe_nasalflare       1   0.0381
## 11 oe_retractions      1   0.0381
```

```
## 12 oe_grunt            1   0.0381
## 13 oh_matfever         1   0.0381
## 14 oh_offliquor        1   0.0381
## 15 pi_gest             0   0
## 16 oe_fontanelle       0   0
## 17 oe_activity         0   0
## 18 oe_colour           0   0
## 19 oe_omphalitis       0   0
## 20 oe_abskin           0   0
## 21 hx_vomit            0   0
## 22 co_prom             0   0
## 23 sepsis              0   0
```

0.1.3 Case-wise missingness

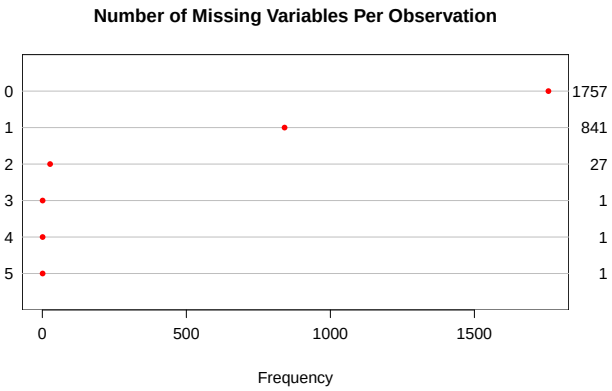
Most participants had no missing data and, among those who did, the majority were only missing values for one predictor (most commonly temperature at admission).

```
## # A tibble: 6 x 3
##   n_miss_in_case n_cases pct_cases
##            <int>   <int>     <dbl>
## 1              0    1757   66.9
## 2              1     841   32.0
## 3              2      27    1.03
## 4              3       1    0.0381
## 5              4       1    0.0381
## 6              5       1    0.0381
```

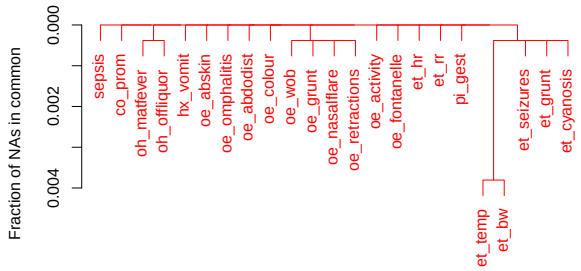



```
##
##  Variables sorted by number of missings:
##        Variable        Count
##         et_temp 0.3097412481
##           et_bw 0.0121765601
##          oe_wob 0.0098934551
##           et_rr 0.0083713851
##           et_hr 0.0011415525
##     oe_abdodist 0.0007610350
##        et_grunt 0.0003805175
```

```
##     et_cyanosis 0.0003805175
##     et_seizures 0.0003805175
##   oe_nasalflare 0.0003805175
##  oe_retractions 0.0003805175
##        oe_grunt 0.0003805175
##     oh_matfever 0.0003805175
##    oh_offliquor 0.0003805175
```



The dendrogram below shows predictors that were commonly missing together.



0.1.4 Relationship between missing temperature and the study outcome

There was no evidence of an association between having a missing value for temperature at admission and the primary outcome of early-onset sepsis:

```
##
##       no  yes
##   0 1596  218
##   1  735   79
```

```
##
## Call:
## glm(formula = sepsis ~ na_temp, family = "binomial", data = dat)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5060  -0.5060  -0.5060  -0.4519   2.1599
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.99076    0.07221 -27.571   <2e-16 ***
## na_temp     -0.23966    0.13867  -1.728   0.0839 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1854.2  on 2627  degrees of freedom
## Residual deviance: 1851.1  on 2626  degrees of freedom
## AIC: 1855.1
##
## Number of Fisher Scoring iterations: 4
```

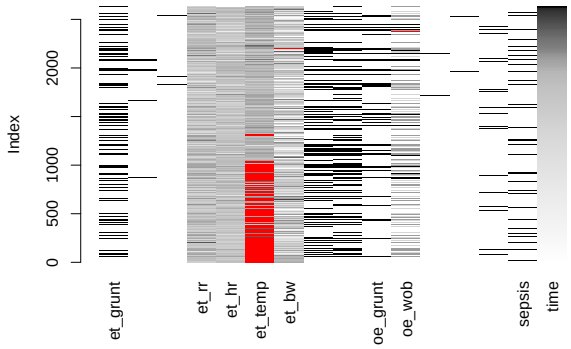
Characteristic	**OR**	**95% CI**	**p-value**
na_temp	0.79	0.60, 1.03	0.084

0.1.5 Relationship between missing temperature and time

Towards the start of the Neotree project, there was a limited number of thermometers available to measure temperature and, therefore, time since the start of the study is a plausible predictor of missingness.

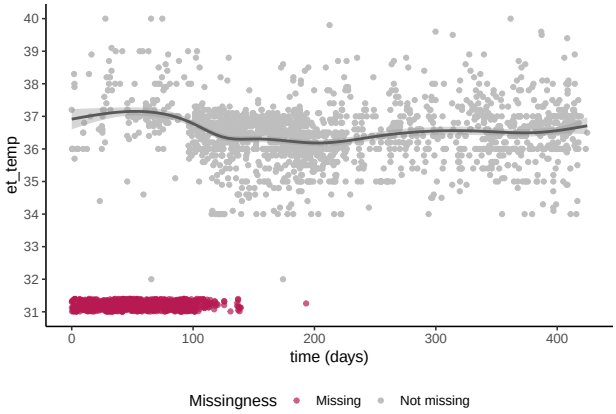
Indeed, most missing values for temperature at admission occurred near the start of data collection. This suggests that temperature was missing at random (MAR) conditional on time since start of the project.

The matrix plot below shows missing values in red, with each participant sorted by their admission date (i.e. time since the start of data collection).



Furthermore, the below figure and a logistic regression analysis demonstrate that time since the start of data collection was a significant predictor of temperature at admission being missing.

Notably, the average recorded temperature was approximately 0.5°C higher during the first 100 days compared to the rest of the data collection period. It is plausible that, during the first 100 days, healthcare workers were more likely to record temperature for 'sicker' babies who were thus more likely to have an elevated temperature. Nevertheless, a wide range of participant characteristics were collected by the Neotree app and were included in the imputation model.




```
##
## Call:
## glm(formula = na_temp ~ time, family = "binomial", data = dat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.64046  -0.32368  -0.02179   0.38527   2.93137
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.455560   0.164777   20.97   <2e-16 ***
## time        -0.040064   0.001724  -23.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3252.9  on 2627  degrees of freedom
## Residual deviance: 1462.6  on 2626  degrees of freedom
## AIC: 1466.6
##
## Number of Fisher Scoring iterations: 7
```

Characteristic	**OR**	**95% CI**	**p-value**
time	0.96	0.96, 0.96	<0.001

0.2 Impute missing values

The imputation model contained all candidate predictors, the outcome of sepsis, and ancillary variables included in the descriptive analysis or that were determined to predict missingness (i.e. time, see above).

Data were assumed to be MAR and 40 imputed datasets were created with 20 iterations. There is no consensus on the optimal number of imputations for multiple imputation, but 40 was chosen based on 33.1% of participants having at least one missing value.

The performance of the imputation model is shown below:

```
## [1] "Imputation method for each variable..."
```

```
##        pi_gest          et_bw    oh_matfever   oh_offliquor        co_prom
##             ""          "pmm"       "logreg"       "logreg"             ""
##      et_grunt          et_rr          et_hr       et_temp    oe_activity
##       "logreg"          "pmm"          "pmm"          "pmm"             ""
## oe_nasalflare oe_retractions       oe_grunt        oe_wob    et_cyanosis
##       "logreg"       "logreg"       "logreg"     "polyreg"       "logreg"
##    et_seizures  oe_fontanelle       oe_colour    oe_abdodist  oe_omphalitis
##       "logreg"             ""             ""       "logreg"             ""
##      oe_abskin       hx_vomit         sepsis           time         pi_sex
##             ""             ""             ""             ""             ""
##         pi_age        outcome
##      "polyreg"             ""
```

```
## [1] "Diagnostic plots..."
```