Additional MIMIC Preprocessing

Rachael Phillips

Contents

1 Previously preprocessed data

We begin by examining the data that was previously preprocessed. The data frame is titled `df` and is provided in the file `data.Rdata`. It is a long file with 22 variables and 1,276 unique patients.

For each 90 min patient's period, there are 60 values for the numerics, corresponding to the *observation period*. The objective is to predict a hypotensive event in the *prediction period* by using only the previous values of the *observation period* after a 30 minute *gap period*.

- Patients are denoted by thier `subject_id`.
- Baseline variables include `bmi`, `age`, `gender`, `sapsi_first`, `sofa_first`, `care_unit`, `admission_type_descr`, `los_icu`, and `los_hospital`.
- Time-varying treatment variables include the following binaries `amine`, `sedation`, `ventilation`.
- Time-varying "numerics" variables include `time_and_date`, `hr`, `abpsys`, `abpdias`, `abpmean`, and `spo2`.
- Time-varying outcome variable is `event` and is defined as an acute hypotensive episode during the patient's prediction period.

Here is how the gap period appears in the data.

```
##         subject_id periode time gender age sapsi_first sofa_first      bmi
## 3021798         20       1   60      F  76          20          8 34.22847
## 3021880         20       2    1      F  76          20          8 34.22847
##         care_unit admission_type_descr los_icu los_hospital amine sedation
## 3021798         4             ELECTIVE       1           5     0        0
## 3021880         4             ELECTIVE       1           5     0        0
##         ventilation       time_and_date hr spo2 abpsys abpdias abpmean
## 3021798           1 2567-03-30 09:46:59 80   97  128.6    59.8    78.8
## 3021880           1 2567-03-30 10:17:59 80   96  109.3    54.6    68.7
##         event
## 3021798     0
## 3021880     1
```

Later, we will examine the data frame `numerics`, which is provided in the file `data_numerics.Rdata`. The `numerics` data does not contain this gap period (i.e., the 30 minutes of missing data is available)

1.1 Missing outcome and erroneous variable values

First let's summarize each variable in `df`.

```
## Skim summary statistics
##  n obs: 7126365
##  n variables: 22
##
## -- Variable type:character ---------------------------------------------------------------------
##              variable missing complete       n min max empty n_unique
##  admission_type_descr  238637  6887728 7126365   6   9     0        4
##                gender       0  7126365 7126365   1   1     0        2
##               periode       0  7126365 7126365   1   5     0     1872
##
## -- Variable type:integer -----------------------------------------------------------------------
##    variable missing complete       n     mean      sd p0  p25   p50   p75
##   care_unit       0  7126365 7126365     3.38    2.17  1    1     4     6
##  subject_id       0  7126365 7126365 13869.53 7640.25 20 7381 13793 20124
##   p100     hist
##      6
##  26711
##
## -- Variable type:numeric -----------------------------------------------------------------------
##      variable missing complete       n   mean      sd     p0    p25    p50
##       abpdias     525  7125840 7126365  58.36   18.7  -22.8   49.6   57.6
##       abpmean     525  7125840 7126365  83.12   23.84   0.1   69.2   78.7
##        abpsys     525  7125840 7126365 119.16   34.41   0    102.3  118.4
##           age       0  7126365 7126365  65.45   17.14   0     55     67
##         amine       0  7126365 7126365   0.51    0.5    0      0      1
##           bmi 2898356  4228009 7126365 117.48 1389.46   1.97  24.23  28.15
##         event       0  7126365 7126365   0.18    0.38   0      0      0
##            hr     525  7125840 7126365  85.27   17.44   0     73.5   84.8
##  los_hospital  238637  6887728 7126365  21.25   19.92   1      8     15
##       los_icu       0  7126365 7126365  11.31   14.74   0      3      6
##   sapsi_first 1117886  6008479 7126365  15.63    5.07   1     12     16
##      sedation       0  7126365 7126365   0.43    0.5    0      0      0
##    sofa_first  435318  6691047 7126365   7.48    4.12   0      4      7
##          spo2     525  7125840 7126365  84.93   32.27   0     94.6   97.1
##          time     525  7125840 7126365  30.5    17.32   1     15.75  30.5
##   ventilation       0  7126365 7126365   0.27    0.44   0      0      0
##     p75     p100     hist
##   67      346.2
##   91.6    361
##  138.5    439.6
##   78      200
##    1        1
##   33.86 22436.29
##    0        1
##   96      259.5
##   28      164
##   15      198
##   19       38
##    1        1
##   10       22
##   99      100
##   45.25    60
##    1        1
##
## -- Variable type:POSIXct -----------------------------------------------------------------------
##       variable missing complete       n        min        max     median
##  time_and_date     525  7125840 7126365 2500-10-27 3501-01-01 3013-08-01
##  n_unique
##   3807554
```

We see that the outcome of interest `abpmean` is missing for 525 observations.

The following erroneous values can be seen from the summary above:

- maximum `bmi` of 22436.3 and minimum `bmi` of 4.04
- maximum `age` of 200

I removed the 525 rows with missing values for `abpmean`.

1.2 Subject IDs representing multiple patients

Next, for every subject, I calculated the number of distinct levels for every baseline covariate. Each subject should have 1 distinct level for each baseline covariate.

```
## # A tibble: 406 x 10
##    subject_id gender   age   bmi care_unit admission_type_- los_icu
##         <int>  <int> <int> <int>     <int>            <int>   <int>
##  1        124      1     3     2         2                2       3
##  2        138      1     2     3         2                2       3
##  3        177      1     2     1         3                1       4
```

```
##  4        214      1     1     3         2               1       2
##  5        283      1     1     2         2               1       2
##  6        328      1     1     2         1               1       2
##  7        377      1     1     1         2               1       3
##  8        408      1     2     1         2               1       3
##  9        507      1     2     2         2               2       2
## 10        638      1     1     1         2               1       2
## # ... with 396 more rows, and 3 more variables: los_hospital <int>,
## #   sapsi_first <int>, sofa_first <int>
```

There are 406 patients with more than one distinct level for each baseline covariate. How does this appear in the data?

```
##   subject_id periode time gender age sapsi_first sofa_first      bmi
## 1        124       1    1      M  71           8          2       NA
## 2        124       1    1      M  75          NA         NA       NA
## 3        124       1    1      M  70          11          2 22.13825
## 4        124       1    2      M  71           8          2       NA
## 5        124       1    2      M  75          NA         NA       NA
## 6        124       1    2      M  70          11          2 22.13825
##   care_unit admission_type_descr los_icu los_hospital amine sedation
## 1      MICU            EMERGENCY       7           7     0        0
## 2      CSRU                 <NA>       1          NA     0        0
## 3      CSRU            EMERGENCY       4          21     0        0
## 4      MICU            EMERGENCY       7           7     0        0
## 5      CSRU                 <NA>       1          NA     0        0
## 6      CSRU            EMERGENCY       4          21     0        0
##   ventilation      time_and_date   hr spo2 abpsys abpdias abpmean event
## 1           0 3297-08-03 13:57:12 51.1 98.0    0.0     0.0    41.2     0
## 2           0 3297-08-03 13:57:12 51.1 98.0    0.0     0.0    41.2     0
## 3           0 3297-08-03 13:57:12 51.1 98.0    0.0     0.0    41.2     0
## 4           0 3297-08-03 13:58:12 51.4 97.5  102.5    55.9    70.9     0
## 5           0 3297-08-03 13:58:12 51.4 97.5  102.5    55.9    70.9     0
## 6           0 3297-08-03 13:58:12 51.4 97.5  102.5    55.9    70.9     0
```

It seems like this subject id is representing more than one patient. However, the outcome values are the same across these seemingly different subjects. Because I do not know which covariate information corresponds to the outcome measurements, I removed all subject id's with multiple baseline covariate values.

After removing these 406 subject ids, which appear to represent multiple patients, 870 subjects remained.

1.3 Outcome measurement error

What does it mean when `abpmean` has a value, but `abpsys` and `abpdias` are both zero? See below.

```
##         subject_id periode time gender age sapsi_first sofa_first      bmi
## 3021795         20       1   27      F  76          20          8 34.22847
## 3021783         20       1   28      F  76          20          8 34.22847
## 3021771         20       1   29      F  76          20          8 34.22847
## 3021925         20       2    9      F  76          20          8 34.22847
## 3021982         20       3   42      F  76          20          8 34.22847
##         care_unit admission_type_descr los_icu los_hospital amine sedation
## 3021795      CSRU            ELECTIVE       1           5     0        0
## 3021783      CSRU            ELECTIVE       1           5     0        0
## 3021771      CSRU            ELECTIVE       1           5     0        0
## 3021925      CSRU            ELECTIVE       1           5     0        0
## 3021982      CSRU            ELECTIVE       1           5     0        0
##         ventilation       time_and_date hr  spo2 abpsys abpdias abpmean
## 3021795           1 2567-03-30 09:13:59 80  99.9      0       0   224.5
## 3021783           1 2567-03-30 09:14:59 80  99.8      0       0   169.9
## 3021771           1 2567-03-30 09:15:59 80 100.0      0       0   142.1
## 3021925           1 2567-03-30 10:25:59 80  95.0      0       0    57.8
## 3021982           1 2567-03-30 12:28:59 80 100.0      0       0   134.4
##         event
## 3021795     0
## 3021783     0
## 3021771     0
## 3021925     1
## 3021982     0
```

There are 65,022 rows where this is the case. Here's how this looks in the data.

```
##         subject_id periode time gender age sapsi_first sofa_first      bmi
## 3021802         20       1   25      F  76          20          8 34.22847
## 3021794         20       1   26      F  76          20          8 34.22847
## 3021795         20       1   27      F  76          20          8 34.22847
## 3021783         20       1   28      F  76          20          8 34.22847
## 3021771         20       1   29      F  76          20          8 34.22847
```

```
## 3021772         20       1   30      F  76          20          8 34.22847
##         care_unit admission_type_descr los_icu los_hospital amine sedation
## 3021802      CSRU             ELECTIVE       1           5     0        0
## 3021794      CSRU             ELECTIVE       1           5     0        0
## 3021795      CSRU             ELECTIVE       1           5     0        0
## 3021783      CSRU             ELECTIVE       1           5     0        0
## 3021771      CSRU             ELECTIVE       1           5     0        0
## 3021772      CSRU             ELECTIVE       1           5     0        0
##         ventilation      time_and_date hr  spo2 abpsys abpdias abpmean
## 3021802           1 2567-03-30 09:11:59 80 100.0  113.8    55.6    72.3
## 3021794           1 2567-03-30 09:12:59 80 100.0  115.7    53.8    73.2
## 3021795           1 2567-03-30 09:13:59 80  99.9    0.0     0.0   224.5
## 3021783           1 2567-03-30 09:14:59 80  99.8    0.0     0.0   169.9
## 3021771           1 2567-03-30 09:15:59 80 100.0    0.0     0.0   142.1
## 3021772           1 2567-03-30 09:16:59 80 100.0  122.0    59.4    77.4
##         event
## 3021802     0
## 3021794     0
## 3021795     0
## 3021783     0
## 3021771     0
## 3021772     0
```

The above `abpmean` values (with `abpsys` and `abpdias` both zero) seem to be inconsistent with the other `abpmean` readings.

However, in other cases (like below), these odd `abpmean` values are consistent with the normal `abpmean` readings.

```
##         subject_id periode time gender age sapsi_first sofa_first      bmi
## 6654207         79       1   15      M  52           2          1 24.42046
## 6654208         79       1   16      M  52           2          1 24.42046
## 6654209         79       1   17      M  52           2          1 24.42046
## 6654210         79       1   18      M  52           2          1 24.42046
## 6654211         79       1   19      M  52           2          1 24.42046
## 6654226         79       1   20      M  52           2          1 24.42046
## 6654227         79       1   21      M  52           2          1 24.42046
## 6654228         79       1   22      M  52           2          1 24.42046
## 6654239         79       1   23      M  52           2          1 24.42046
##         care_unit admission_type_descr los_icu los_hospital amine sedation
## 6654207      CSRU            EMERGENCY       2           4     0        0
## 6654208      CSRU            EMERGENCY       2           4     0        0
## 6654209      CSRU            EMERGENCY       2           4     0        0
## 6654210      CSRU            EMERGENCY       2           4     0        0
## 6654211      CSRU            EMERGENCY       2           4     0        0
## 6654226      CSRU            EMERGENCY       2           4     0        0
## 6654227      CSRU            EMERGENCY       2           4     0        0
## 6654228      CSRU            EMERGENCY       2           4     0        0
## 6654239      CSRU            EMERGENCY       2           4     0        0
##         ventilation      time_and_date    hr  spo2 abpsys abpdias abpmean
## 6654207           0 2756-08-13 16:45:00 102.8  98.4    0.0     0.0    83.2
## 6654208           0 2756-08-13 16:46:00 102.7  99.0   97.0    71.2    82.1
## 6654209           0 2756-08-13 16:47:00  98.3  99.5    0.0     0.0    78.4
## 6654210           0 2756-08-13 16:48:00  99.3 100.0    0.0     0.0    79.4
## 6654211           0 2756-08-13 16:49:00 100.3  99.3    0.0     0.0    80.1
## 6654226           0 2756-08-13 16:50:00 100.8  99.1    0.0     0.0    77.1
## 6654227           0 2756-08-13 16:51:00 101.7  98.0   94.2    68.5    79.2
## 6654228           0 2756-08-13 16:52:00 104.4  95.2    0.0     0.0    79.0
## 6654239           0 2756-08-13 16:53:00 101.5  84.4    0.0     0.0    80.9
##         event
## 6654207     0
## 6654208     0
## 6654209     0
## 6654210     0
## 6654211     0
## 6654226     0
## 6654227     0
## 6654228     0
## 6654239     0
```

To solve this issue, I calculated subject specific `abpmean` outlier thresholds as a way to determine if these odd `abpmean` values are outliers or not.

It's important to note that the subject specific distributions of `abpmean` values did consider the missing data in the gap period. Section 2.4.1 resolves this issue.

For 2 subjects (ids 25373 and 26209), *all* of the outcome measurements contain zeros for `abpsys` and `abpdias` and values for `abpmean`, so the outlier thresholds could not be calculated. These two subjects were removed from the data. Now we can see the outliers.

```
##   subject_id periode time gender age sapsi_first sofa_first      bmi
## 1         20       1   27      F  76          20          8 34.22847
## 2         20       1   28      F  76          20          8 34.22847
## 3         20       1   29      F  76          20          8 34.22847
## 4         20       2    9      F  76          20          8 34.22847
## 5         20       3   42      F  76          20          8 34.22847
##   care_unit admission_type_descr los_icu los_hospital amine sedation
## 1      CSRU             ELECTIVE       1           5     0        0
## 2      CSRU             ELECTIVE       1           5     0        0
## 3      CSRU             ELECTIVE       1           5     0        0
## 4      CSRU             ELECTIVE       1           5     0        0
## 5      CSRU             ELECTIVE       1           5     0        0
##   ventilation       time_and_date hr  spo2 abpsys abpdias abpmean event
## 1           1 2567-03-30 09:13:59 80  99.9      0       0   224.5     0
## 2           1 2567-03-30 09:14:59 80  99.8      0       0   169.9     0
## 3           1 2567-03-30 09:15:59 80 100.0      0       0   142.1     0
## 4           1 2567-03-30 10:25:59 80  95.0      0       0    57.8     1
## 5           1 2567-03-30 12:28:59 80 100.0      0       0   134.4     0
##   outlier
## 1       1
## 2       1
## 3       1
## 4       0
## 5       1
```

Of the 65,022 rows with this odd outcome measurement, 31,209 rows were deemed outliers and were subsequently removed.

Now 868 subjects remain in the data.

1.4 Missing baseline characteristics

There are still missing values in the data.

```
##          subject_id             periode                time
##                   0                   0                   0
##              gender                 age         sapsi_first
##                   0                   0               55185
##          sofa_first                 bmi           care_unit
##               44167              678023                   0
## admission_type_descr            los_icu        los_hospital
##              13671                   0               13671
##               amine            sedation         ventilation
##                   0                   0                   0
##       time_and_date                  hr                spo2
##                   0                   0                   0
##              abpsys             abpdias             abpmean
##                   0                   0                   0
##               event
##                   0
```

Removing all NA values would lead to the omission of 308 subjects, but 258 of these subjects were only missing bmi and no other value. Many subjects had very odd bmi values such as 4, 5, 9, 69, 77, and 22436. Because of this oddity, I only removed subjects that were missing values for other covariates.

818 subjects remain in the data.

```
##          subject_id             periode                time
##                   0                   0                   0
##              gender                 age         sapsi_first
##                   0                   0                   0
##          sofa_first                 bmi           care_unit
##                   0              647184                   0
## admission_type_descr            los_icu        los_hospital
##                   0                   0                   0
##               amine            sedation         ventilation
##                   0                   0                   0
##       time_and_date                  hr                spo2
##                   0                   0                   0
##              abpsys             abpdias             abpmean
##                   0                   0                   0
##               event
##                   0
```

1.5 Insufficient patient data

As described in the beginning, the objective is to predict a hypotensive event in the prediction period (i.e., after 90 minutes) using only the values in the observation period (i.e., the first 60 minutes) after a 30 minute gap period.

We need each patient to have data in the prediction period, so we can evaulate the predictions with their actual value. Thus, we need each subject to have data for over 90 minutes to ensure that the data extends into the prediction period.

However, of the 818 remaining subjects, 21 had less than 90 minutes of data available. These subjects were removed from the data.

797 subjects remain in the data.

1.6 Classifying a hypotensive event

The variable `event` is defined as an acute hypotensive episode during the patient's prediction period. In some cases, it appears that a hypotensive event is not classified as one. In other cases, it appears that a non-hypotensive event is misclassified as a hypotensive event. Both scenarios are shown below.

I created a new outcome `Y1` using a function Ivana made. This function classifies an event as hypotensive when the current `abpmean` is less than 62 and at least 5 adjacent *time points* have `abpmean` less than 65. Adjacent time points are typically 1 minute apart. There is an exception for time points that occur 5 minutes before gap periods and 5 minutes after gap periods. These 5 minutes before and after gap periods consider adjacent time points that are separated by 30 minutes.

See Section 2.4.1 for a resolution to this issue.

We can examine outcomes `Y1` and `event` below.

```
##       subject_id periode time gender age sapsi_first sofa_first      bmi
## 86449        906      18    8      M  78          22         14 19.67777
## 86450        906      18    9      M  78          22         14 19.67777
## 86451        906      18   10      M  78          22         14 19.67777
## 86452        906      18   11      M  78          22         14 19.67777
## 86453        906      18   12      M  78          22         14 19.67777
##       care_unit admission_type_descr los_icu los_hospital amine sedation
## 86449      MICU            EMERGENCY      13           14     1        0
## 86450      MICU            EMERGENCY      13           14     1        0
## 86451      MICU            EMERGENCY      13           14     1        0
## 86452      MICU            EMERGENCY      13           14     1        0
## 86453      MICU            EMERGENCY      13           14     1        0
##       ventilation       time_and_date    hr spo2 abpsys abpdias abpmean
## 86449           0 2653-04-21 02:50:25  95.6  100   76.8    47.9    60.0
## 86450           0 2653-04-21 02:51:25 103.0  100   65.7    45.0    53.0
## 86451           0 2653-04-21 02:52:25  99.8  100   70.9    50.8    58.7
## 86452           0 2653-04-21 02:53:25  97.6  100    0.0     0.0    54.1
## 86453           0 2653-04-21 02:54:25  98.2  100   79.0    42.3    58.4
##       event Y1
## 86449     0  1
## 86450     0  1
## 86451     0  1
## 86452     0  1
## 86453     0  1

##    subject_id periode time gender age sapsi_first sofa_first      bmi
## 86         20       2   29      F  76          20          8 34.22847
## 87         20       2   30      F  76          20          8 34.22847
## 88         20       2   31      F  76          20          8 34.22847
## 89         20       2   32      F  76          20          8 34.22847
## 90         20       2   33      F  76          20          8 34.22847
## 91         20       2   34      F  76          20          8 34.22847
## 92         20       2   35      F  76          20          8 34.22847
##    care_unit admission_type_descr los_icu los_hospital amine sedation
## 86      CSRU            ELECTIVE       1           5     0        0
## 87      CSRU            ELECTIVE       1           5     0        0
## 88      CSRU            ELECTIVE       1           5     0        0
## 89      CSRU            ELECTIVE       1           5     0        0
## 90      CSRU            ELECTIVE       1           5     0        0
```

```
## 91      CSRU             ELECTIVE       1           5     0        0
## 92      CSRU             ELECTIVE       1           5     0        0
##    ventilation       time_and_date hr  spo2 abpsys abpdias abpmean event
## 86           1 2567-03-30 10:45:59 80  96.7   95.1    50.2    62.0     1
## 87           1 2567-03-30 10:46:59 80  98.5  118.7    59.8    75.9     1
## 88           1 2567-03-30 10:47:59 80  99.9  140.3    68.1    89.1     1
## 89           1 2567-03-30 10:48:59 80 100.0  154.2    73.8    98.6     1
## 90           1 2567-03-30 10:49:59 80 100.0  161.2    76.9   104.0     1
## 91           1 2567-03-30 10:50:59 80 100.0  164.4    78.1   106.9     1
## 92           1 2567-03-30 10:51:59 80 100.0  155.6    73.2   100.0     1
##    Y1
## 86  1
## 87  0
## 88  0
## 89  0
## 90  0
## 91  0
## 92  0
```

1.7 Summary

This updated data frame with 797 subjects is named `mimic` and does not contain

- NA values for any variable except `bmi`,
- outcome measurement error outliers,
- single subject id's which represent multiple patients,
- or patients with less than 90 minutes of data.

This `mimic` data frame does contain

- the gap period of 30 minutes after every hour of data;
- a new hypotensive event outcome `Y1` which classifies a hypotensive event as one when the current `abpmean` is less than 65 and at least 5 adjacent time points have `abpmean` less than 65;
- and subject's with erroneous baseline characteristic measurements, including
  - `subject_id` = 20936 with `age` of 200,
  - and many subjects with odd `bmi` values such as 4, 5, 9, 77, 69, 22436.

The `mimic` data is used in the Section 2. If you would like to work with data with gap periods, then use the `mimic_gap` data, which is presented in Section 2.4.1.

2 Data numerics

There is another a data frame `numerics` saved in `data_numerics.Rdata`. This data frame contains the outcome data for each subject with no gap periods.

2.1 Omitting 30 minute gap periods

I merged the 30 minute gap period into the `mimic` data that was created in the previous section. I filled in this missing 30 minutes of data because it is preferred for the simulation, since it provides a more clear representation of the patient data.

The only variables in `numerics` data frame are `subject_id`, `time_and_date`, `hr`, `abpsys`, `abpdias`, `abpmean`, and `spo2`.

Before merging the data, I removed any erroneous/outlier outcome measurements according to the method described in Section 1.3.

Here's how the data looks after merging.

```
##    subject_id       time_and_date   hr abpsys abpdias abpmean spo2 periode
## 58      10013 2564-11-02 03:24:18 93.4  105.5    34.8    54.8 92.9       1
## 59      10013 2564-11-02 03:25:18 95.7  106.4    35.3    55.8 92.6       1
## 60      10013 2564-11-02 03:26:18 94.0  104.8    34.7    54.6 92.8      NA
## 61      10013 2564-11-02 03:27:18 93.7  104.1    34.3    53.9 92.0      NA
##    time gender age sapsi_first sofa_first      bmi care_unit
## 58   59      F  87          15          8 34.85214      MICU
## 59   60      F  87          15          8 34.85214      MICU
## 60   NA   <NA>  NA          NA         NA       NA      <NA>
## 61   NA   <NA>  NA          NA         NA       NA      <NA>
##    admission_type_descr los_icu los_hospital amine sedation ventilation
## 58            EMERGENCY       3           3     1        0           0
## 59            EMERGENCY       3           3     1        0           0
## 60                 <NA>      NA          NA  <NA>     <NA>        <NA>
## 61                 <NA>      NA          NA  <NA>     <NA>        <NA>
##    event Y1
## 58     1  1
## 59     1  1
## 60  <NA> NA
## 61  <NA> NA
```

2.2 Fill in missing values

We can see that missing values need to be filled in for

1. the time-varying treatments
2. and the baseline covariates.

I filled in the missing baseline covariate information with `tidyr::fill`. This function fills missing values using the previous entry, so we assume that the 30 minute gap period had the same baseline covariate values as the minute *before* this gap started. This procedure is surely reasonable for the baseline covariates, but probably not for the time-varying treatment, which is why we didn't fill in that information. Let's see how the data shown above looks after filling in this information.

```
##    subject_id       time_and_date hr abpsys abpdias abpmean spo2 periode
## 58         20 2567-03-30 09:47:59 80  129.3    60.1    79.2   97      NA
## 59         20 2567-03-30 09:48:59 80  128.7    59.9    78.8   97      NA
## 60         20 2567-03-30 09:49:59 80  129.0    60.1    78.9   97      NA
## 61         20 2567-03-30 09:50:59 80  129.0    60.1    78.9   97      NA
##    time gender age sapsi_first sofa_first      bmi care_unit
## 58   NA      F  76          20          8 34.22847      CSRU
## 59   NA      F  76          20          8 34.22847      CSRU
## 60   NA      F  76          20          8 34.22847      CSRU
## 61   NA      F  76          20          8 34.22847      CSRU
##    admission_type_descr los_icu los_hospital amine sedation ventilation
## 58             ELECTIVE       1           5  <NA>     <NA>        <NA>
## 59             ELECTIVE       1           5  <NA>     <NA>        <NA>
## 60             ELECTIVE       1           5  <NA>     <NA>        <NA>
## 61             ELECTIVE       1           5  <NA>     <NA>        <NA>
##    event Y1
## 58  <NA> NA
## 59  <NA> NA
## 60  <NA> NA
## 61  <NA> NA
```

There are some cases when missing values needed to be filled in using a *later* entry, because there are no prior entries without NA. This is the case for 173 subjects. Here's an example of how it looks.

```
##       subject_id       time_and_date    hr abpsys abpdias abpmean  spo2
## 22919        439 3242-12-28 15:07:01 105.8  166.6    72.2   110.0  99.9
## 22920        439 3242-12-28 15:08:01 106.7  166.4    71.2   109.5 100.0
## 22921        439 3242-12-28 15:09:01 107.1  166.5    71.6   109.7 100.0
##       periode time gender age sapsi_first sofa_first bmi care_unit
## 22919      NA   NA   <NA>  NA          NA         NA  NA      <NA>
## 22920      NA   NA   <NA>  NA          NA         NA  NA      <NA>
## 22921       2    1      F  83          33         18  NA       CCU
##       admission_type_descr los_icu los_hospital amine sedation ventilation
## 22919                 <NA>      NA          NA  <NA>     <NA>        <NA>
## 22920                 <NA>      NA          NA  <NA>     <NA>        <NA>
## 22921            EMERGENCY       7          15     0        0           0
##       event Y1
## 22919  <NA> NA
## 22920  <NA> NA
## 22921     0  0
```

The column `periode` refers to the "number of the patient's period". For these 173 subjects, the minimum `periode` is always 2 or higher. Thus, the merge filled in something like `periode` 1 for these subjects.

There was probably a reason for omitting `periode` 1 for these subjects, maybe the time-varying treatment data was unavailable. I can remove `periode` 1 from these subjects later if need be.

Something else I noticed about `periode` when examining this issue. Below I listed all of the unique `periode` available for the subject above, `subject_id` = 439.

```
##   [1]  NA   2   3   4   5   6   7   8  12  13  14  16  17  18  19  20  21
##  [18]  22  23  24  25  26  27  28  29  30  31  33  34  36  37  38  39  40
##  [35]  41  42  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58
##  [52]  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  76
##  [69]  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93
##  [86]  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
```

Interestingly, this subject is missing some of the `periode` (e.g., 9-11). The data that would have been `periode` 9-11 is available in `numerics`, but this information not included in the preprocessed data `df`. Perhaps because the time-varying treatment data was not available.

2.3 Classifying a hypotensive event

Like before, I created a new outcome `Y1` to classify hypotensive events. The issue with adjacent time points occasionally being 30 minutes apart is not present in this scenario, since there is no gap period.

2.4 Summary

This updated data frame with the full data still contains 797 subjects and is named `mimic_nogap`. This updated data does not contain

- NA values for any baseline characteristic variable except `bmi`,
- outcome measurement error outliers,
- single subject id's which represent multiple patients,
- patients with less than 90 minutes of data,
- or the gap period of 30 minutes after every hour of data.

This `mimic_nogap` data frame does contain

- a new hypotensive event outcome `Y1` which classifies a hypotensive event as one when the current `abpmean` is less than 65 and at least 5 adjacent time points have `abpmean` less than 65;
- subject's with erroneous baseline characteristic measurements, including
  - `subject_id` = 20936 with `age` of 200,
  - and many subjects with odd `bmi` values such as 4, 5, 9, 77, 69, 22436;
- NA values during the merged in 30 minute gap period for
  - time-varying treatments `amine sedation ventilation`,
  - and `periode time event`.

The `mimic_nogap` data is saved in `mimic_nogap.Rdata`.

2.4.1 Important Note

Because `mimic_nogap` considered the full data, two of the preprocessing steps in `mimic_nogap` are more reliable than those in `mimic`.

1. The new hypotensive event outcome `Y1` relies on adjacent time points for classification. Since `mimic_nogap` does not contain 30-minute gaps with no data, the adjacent time points are closer together in `mimic_nogap`.

2. There are many instances when `abpmean` has a value, but `abpsys` and `abpdias` are both zero. Sometimes these oddities appear to be consistent with the patients "normal" readings (i.e., readings which have non-zero values for `abpmean`, `abpsys` and `abpdias`). In other cases, these oddities are very different from the patients normal readings. I remove the inconsistent oddities by calculating subject specific `abpmean` outlier thresholds, which requires calculating subject specific IQRs. These IQRs are more accurate in `mimic_nogap`, since the full data is used.

For these reasons, I used `mimic_nogap` to recreate the same gaps that are in `mimic`. I called this data frame `mimic_gap` and saved it in `mimic_gap.Rdata`.

I recommend using the `mimic_gap` data instead of `mimic` data.