Burn Data EDA

Overview

Two studies were conducted:

1) TT - Transfusion Trigger (= TRIBE)

- Liberal vs. restrictive transfusion strategy (more or less blood).
- Results already analyzed and showed no signficant effect of treatment.

2) PCR SEPSIS

- Ucing PCR to try to identify people who had bloodstream infections, but faster than traditional methods

Question addressed here:

- These studies included daily measurements of patients vitals such as temperature, heart-rate, blood pressure, and sodium levels.

Can we use this body of patient information to predict infection?

This work is exploratory "proof of concept" to see if there is something here worth more investigation.

Notes

- Infection variables (e.g., blood, urine) refer to type of infection, with primary interest in bloodstream infections. For the TT study I started EDA focusing on the "any" infection variable, i.e. starting at the most general level, and looked specifically at "blood" infection in some cases, the infection type of primary interest.
- I looked at the day before onset, and in most cases restricted futher to a patient's first onset.
- Sepsis is identified differently for burn patients than others. Individual variables such as temperature which are normally used as indicators, are generally elevated or otherwise abnormal due to the burn.
- Important notes from the study authors, Dr. Palmieri and Dr. Tran via Sandy Taylor:
  - Temperature was used as a criteria for culture in both studies. This means that the strength of temperature as an indicator or pre-indicator of infection is biased, since it was used to select who was screened. This may be true of other measurements. From Dr. Tran's email: *Cultures for PCR Sepsis were collected only when indicated at each site. Typically, that would be based on the presence of signs of sepsis such as fever (Temp >39.5C). For PCR Sepsis, we kept it a bit more simple especially for blood cultures. If the patient had a truly pathogenic organism from blood culture, then they were bacteremic, and considered septic if they met other criteria (I believe we used the same "Infection" form as Transfusion Trigger) such as fever, WBC, platelets, etc.*
  - From Dr. Palmieri's email: *In terms of labs to focus on, would use platelets, wbc, sodium, chloride.*
  - From Dr. Tran's email: *Routinely collected labs that may have value in predicting sepsis (more so burn sepsis) would be the usual CBC. Platelet count useful for sepsis severity and somewhat of a late marker for burn sepsis. We're messing with some AI and machine learning work right now with the PCR Sepsis database and platelets remain a strong parameter to look at. Not sure if TRIBE has CBC indices (RDW, MCHC, MCV, etc) which may be of use. Electrolytes including sodium variability may be helpful from the chemistry panel. Respiratory rate, heart rate, etc are also useful. Pretty much the parameters from Dr. Greenhalgh's consensus guidelines (JBCR 2007?).*

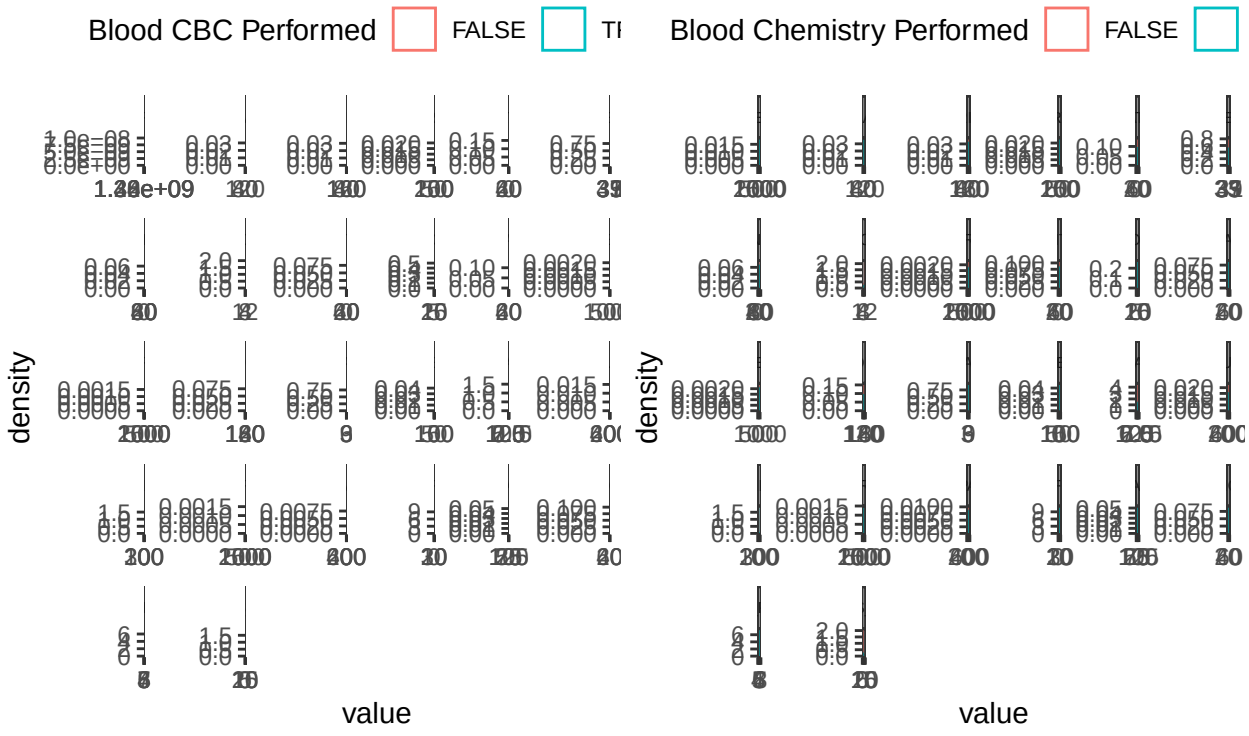
TT Study EDA

Question 0: Describe who was screened for blood infection and when

There are three variables indicating time of blood-level screening in the data, in addition to the vital screening variable (V_TIME_PERFORMED). They refer to blood CBC (V_TIME_PERFORMED_1), blood chemistry (V_TIME_PERFORMED_2), and blood gas (V_TIME_PERFORMED_3) respectively, according to the variable descriptions.

Almost all entries have vitals (V_TIME_PERFORMED, 33 NA), and most have blood CBC (V_TIME_PERFORMED_1, 2186 NA) and blood chemistry (V_TIME_PERFORMED_2, 3542 NA). Only about one third have V_TIME_PERFORMED_3 (9111 NA) out of 14852 entries.

Only two entries have Blood infection reported when neither Blood CBC and Chemisty were performed, but in general (the other 120), this seems to be a precursor for detection.



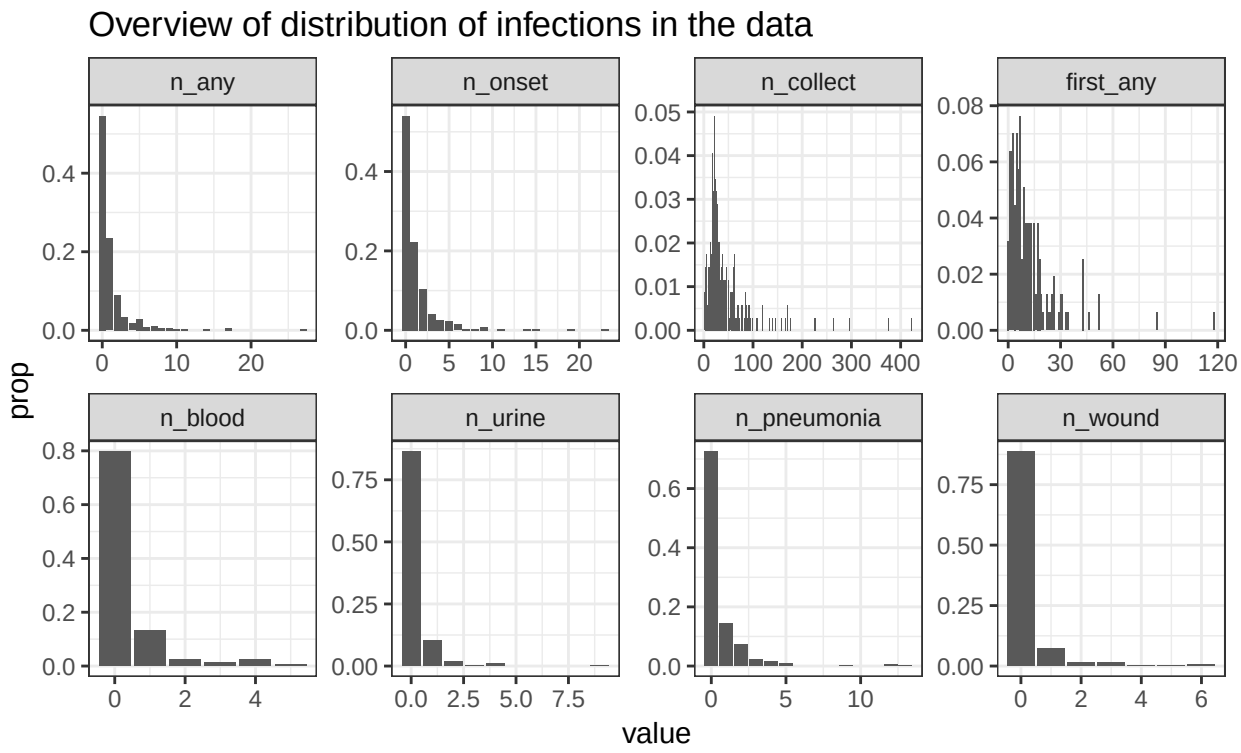
Question 1: Describe the type, frequency and patterns of infections in study participants

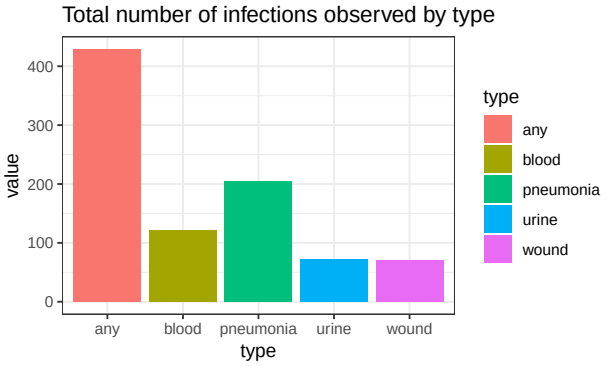
Observations:

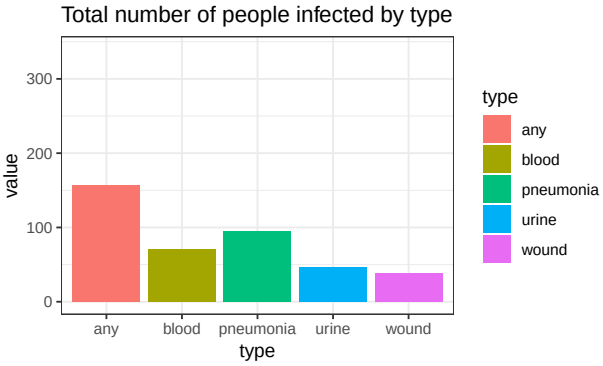
- The majority of individuals in the study had no infection and about a quarter had more than one.
- We'll need to distinguish the first onset of infection from subsequent onsets. Dr. Taylor suggested (12/6 meeting) that data about subsequent infectons may be less reliable than data about the first one.
- The number of data collections for each individual ranges quite a bit, but most had between 0 and 100 collections.
- Only about 20 percent of infections were specifically blood, urine, pneumonia or wound infection. This means that there is limited data to model the first onset of blood infection, which is the outcome we

are most intersted in.

- There is generally low corellation between different infection types.

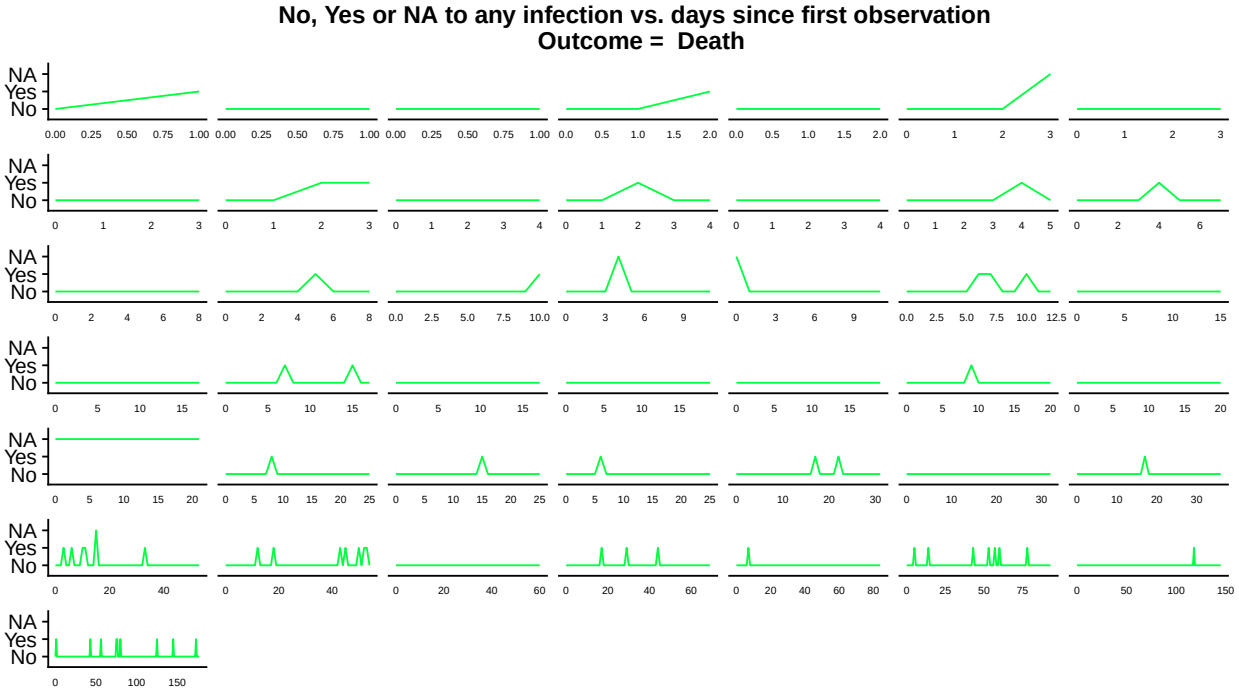


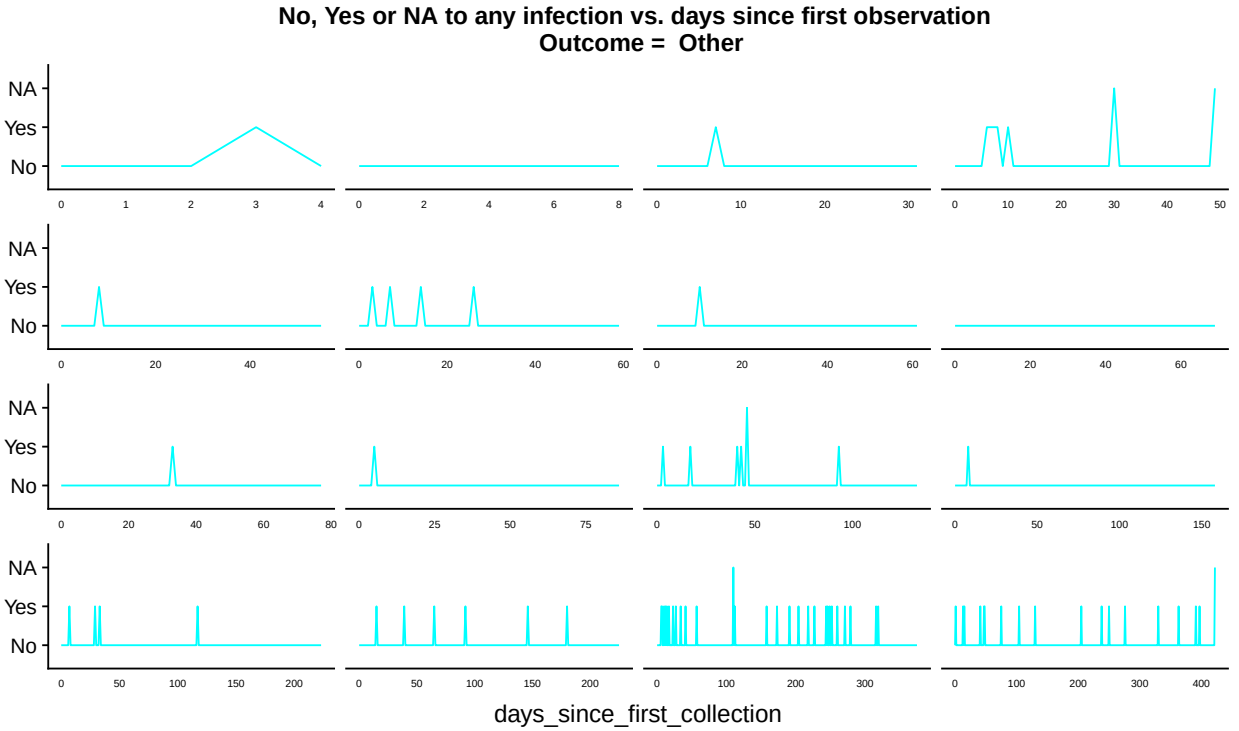


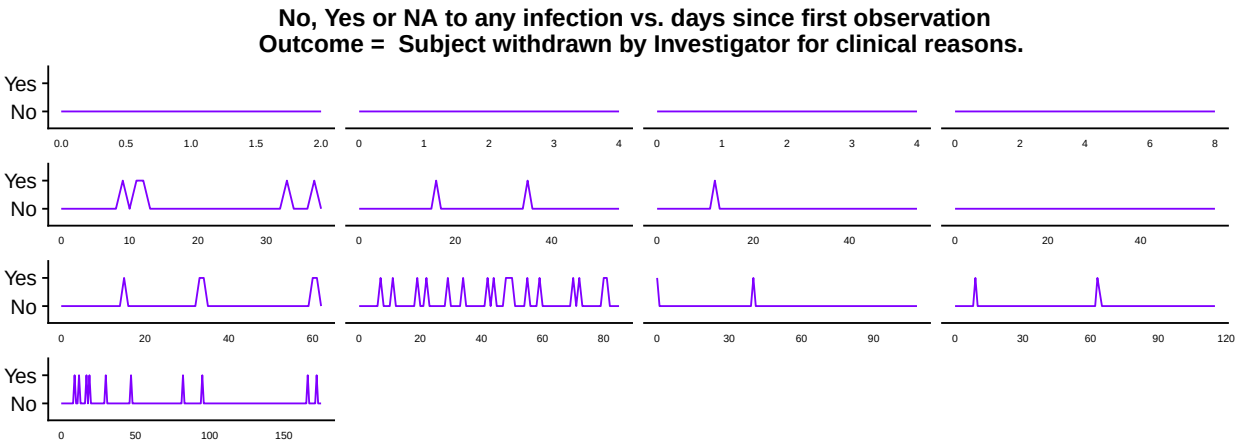


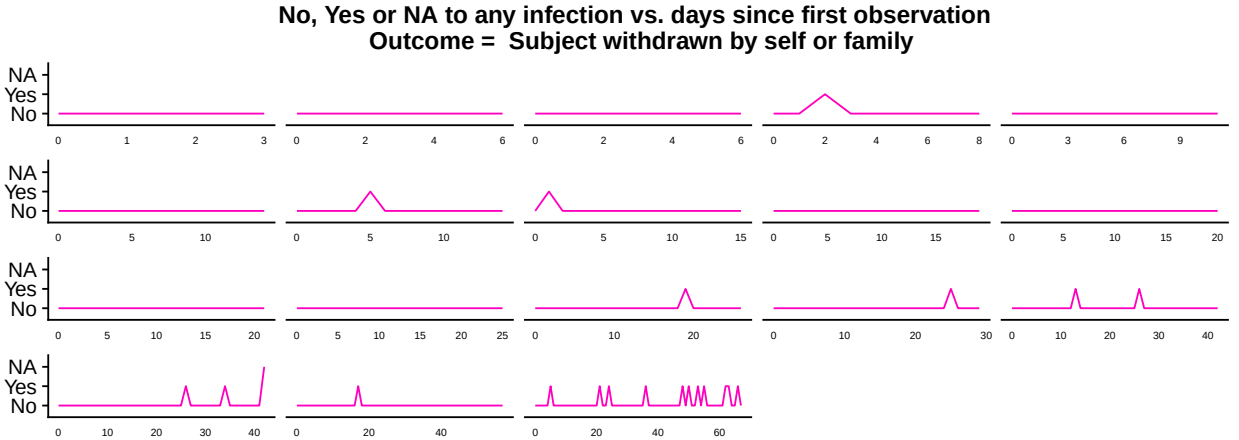

```
#Correlation of having such an infection
round(cor(TT_per[,c("n_blood","n_urine","n_wound", "n_pneumonia", "n_any")] > 0, use = "pairwise.comple

#Correlation of number of such infections
round(cor(TT_per[,c("n_blood","n_urine","n_wound", "n_pneumonia", "n_any")], use = "pairwise.complete.ob
```

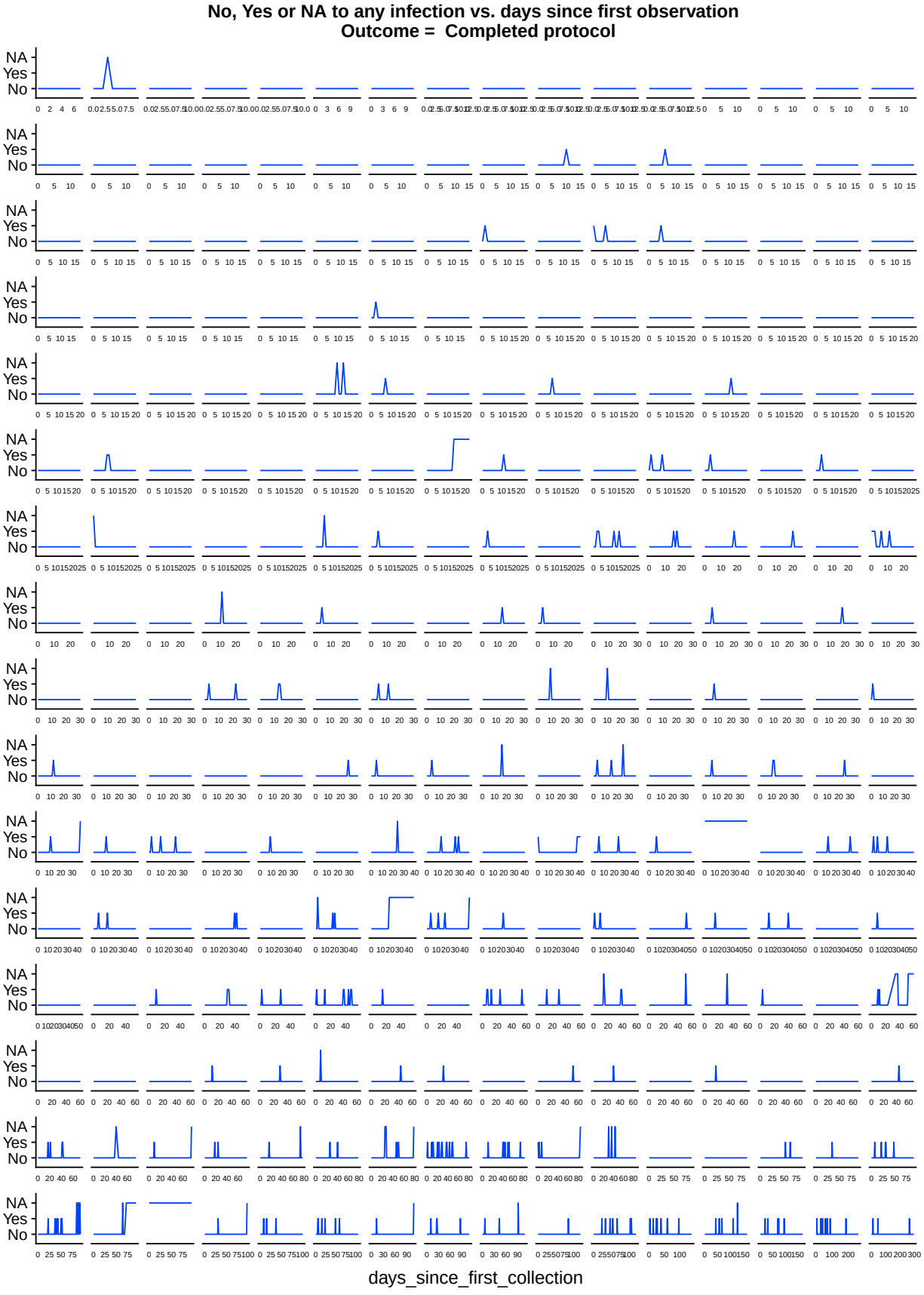
The next five plots show all individuals in the study, grouped by their study outcome. From the day of their first collection, their No/Yes/NA infection status is tracked. Within each outcome group, individuals are sorted by their number of days observed.






No, Yes or NA to any infection vs. days since first observation
Outcome = Subject withdrawn by Investigator for clinical reasons.
Yes
No
days_since_first_collection

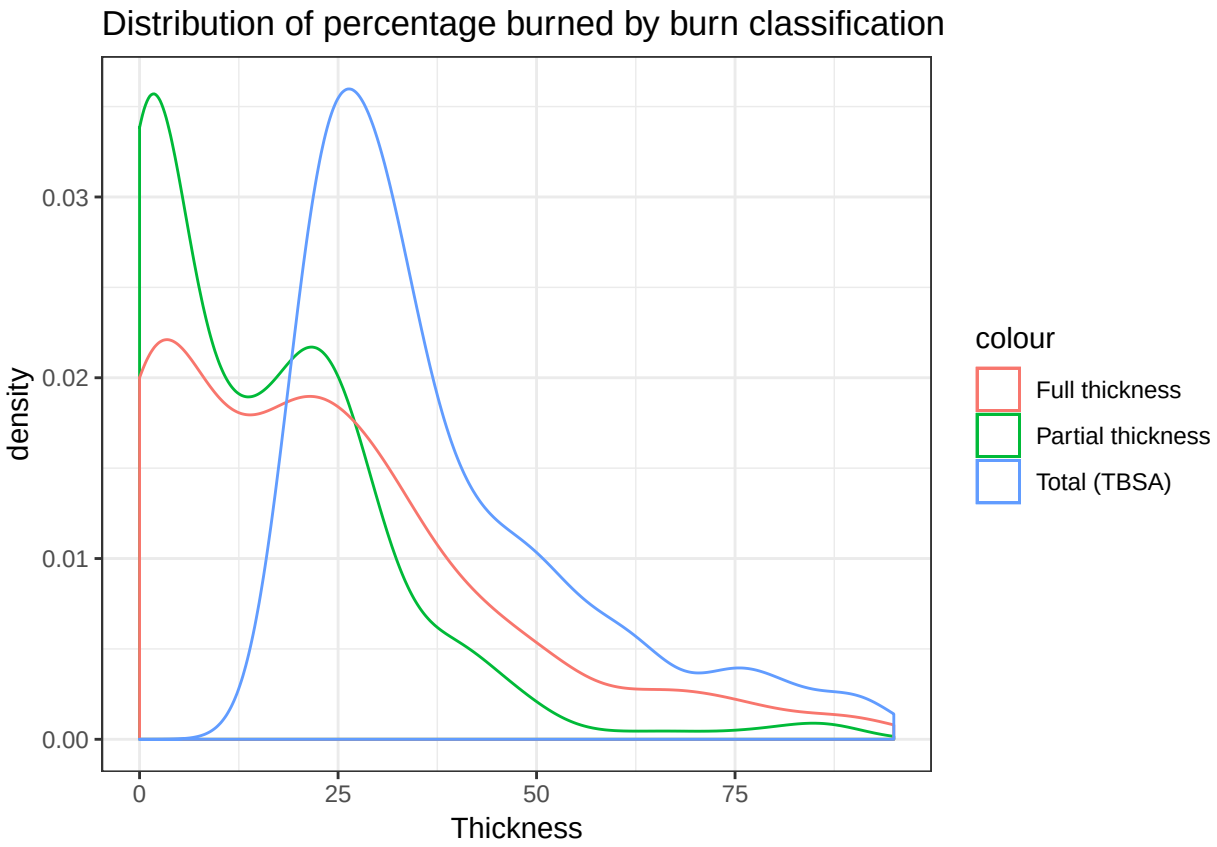

No, Yes or NA to any infection vs. days since first observation
Outcome = Subject withdrawn by self or family
NA
Yes
No
days_since_first_collection

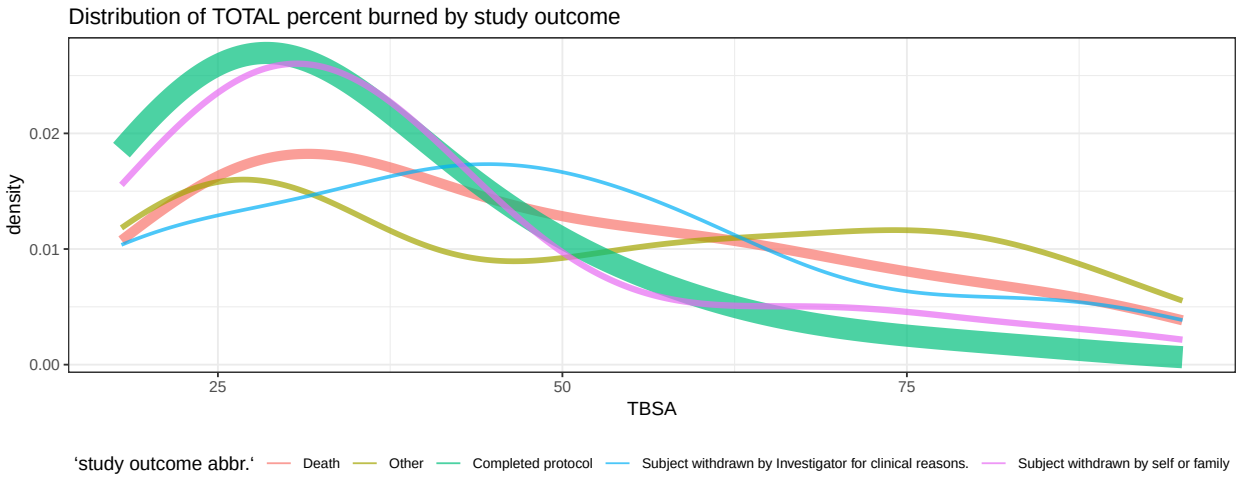

No, Yes or NA to any infection vs. days since first observation
Outcome = Completed protocol
NA
Yes
No
days_since_first_collection

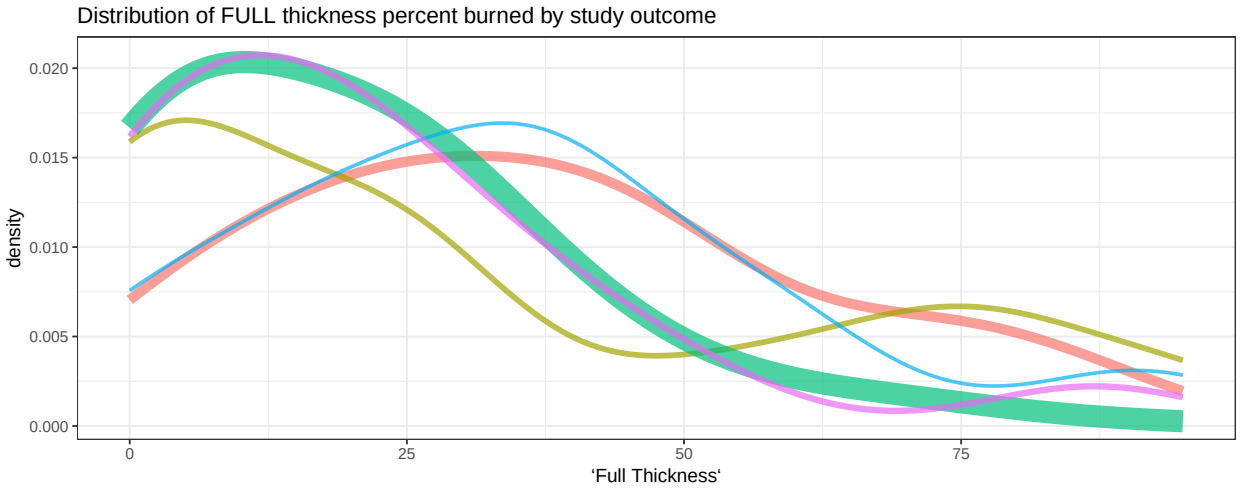
Question 2: Describe the distribution of patient burn severity and connection to outcomes

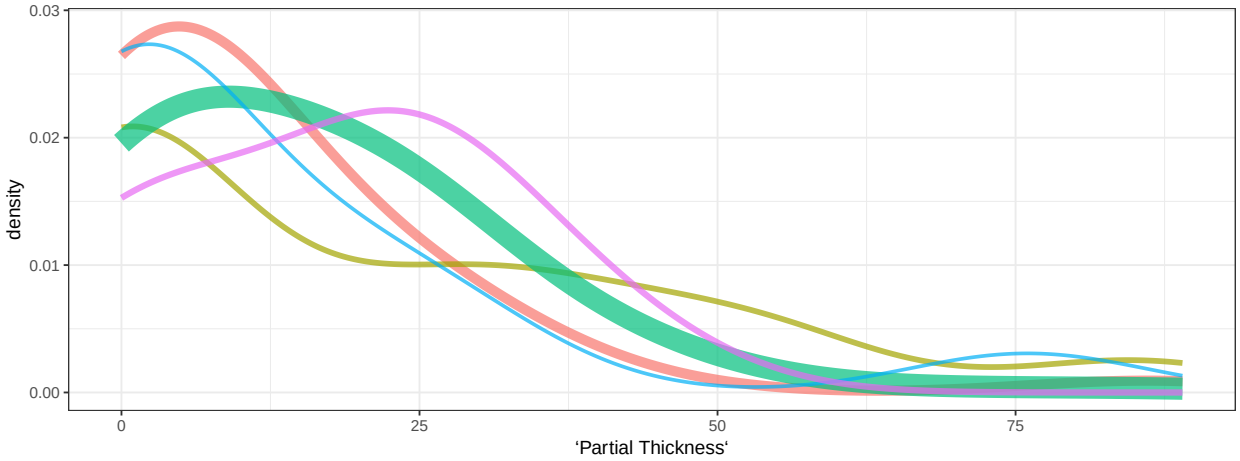
Observations:

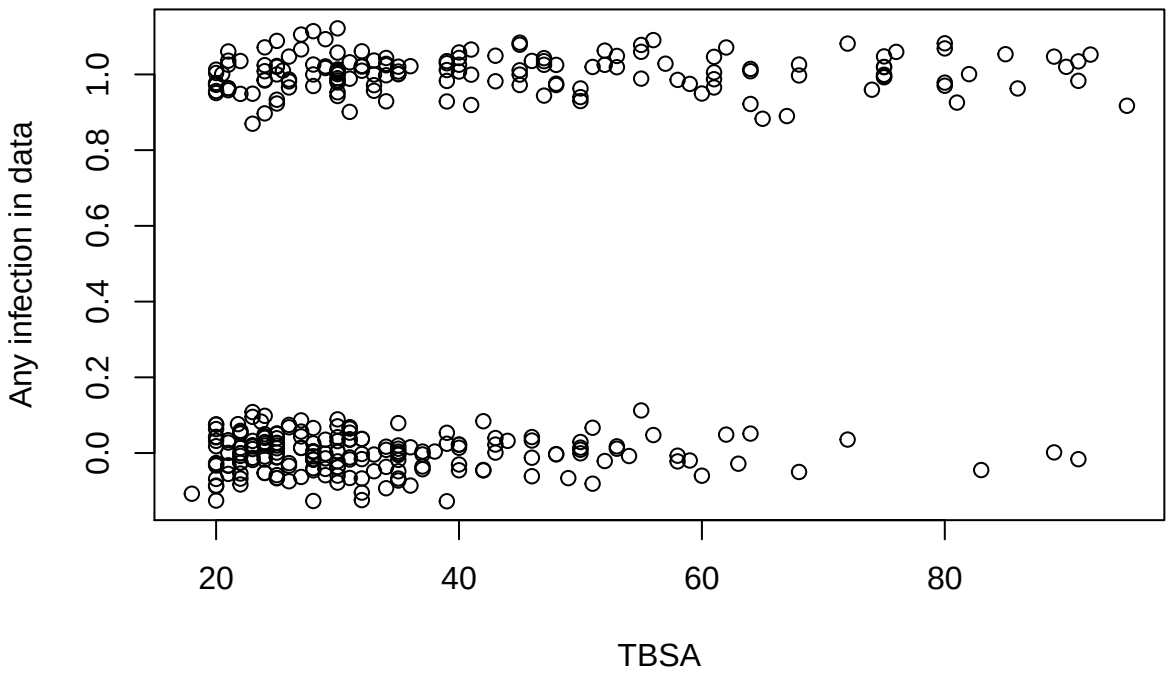
- Most have between 20 and 50 percent of their body burned. Range of TBSA in the study is 18 - 95.
- Overall those with higher percentage burned are more likely for their outcome to be death or withdrawal for clinical reasons as opposed to completing the protocol or withdrawal by self or family. Those with "other" outcomes are also more likely to have higher burn percentages. The thicknesses of the lines in the plots by study outcome correspond to the number of people with that outcome.
- Although most people had no infection, those with TBSA above 60 are more likely than not to have at least one infection.




Distribution of TOTAL percent burned by study outcome
density
0.02
0.01
0.00
25
50
75
TBSA
'study outcome abbr.'
Death
Other
Completed protocol
Subject withdrawn by Investigator for clinical reasons.
Subject withdrawn by self or family


Distribution of FULL thickness percent burned by study outcome
density
0.020
0.015
0.010
0.005
0.000
0
25
50
75
'Full Thickness'


Distribution of PARTIAL thickness percent burned by study outcome
density
0.03
0.02
0.01
0.00
0
25
50
75
'Partial Thickness'



Question 3: Individual vital signs over time

Observations

- It's most common for first infection to occur with in a week of the first data collection, but there is considerable spread after that.
- Most individuals have their first collection within one day of admittance.
- Relatively few of the vitals have strong correlations, and about half of the ones that do are different types of blood pressure. This is true for aggregated data for infection patients/days and non. Individual-level data may be more telling.

When does the first infection occur?

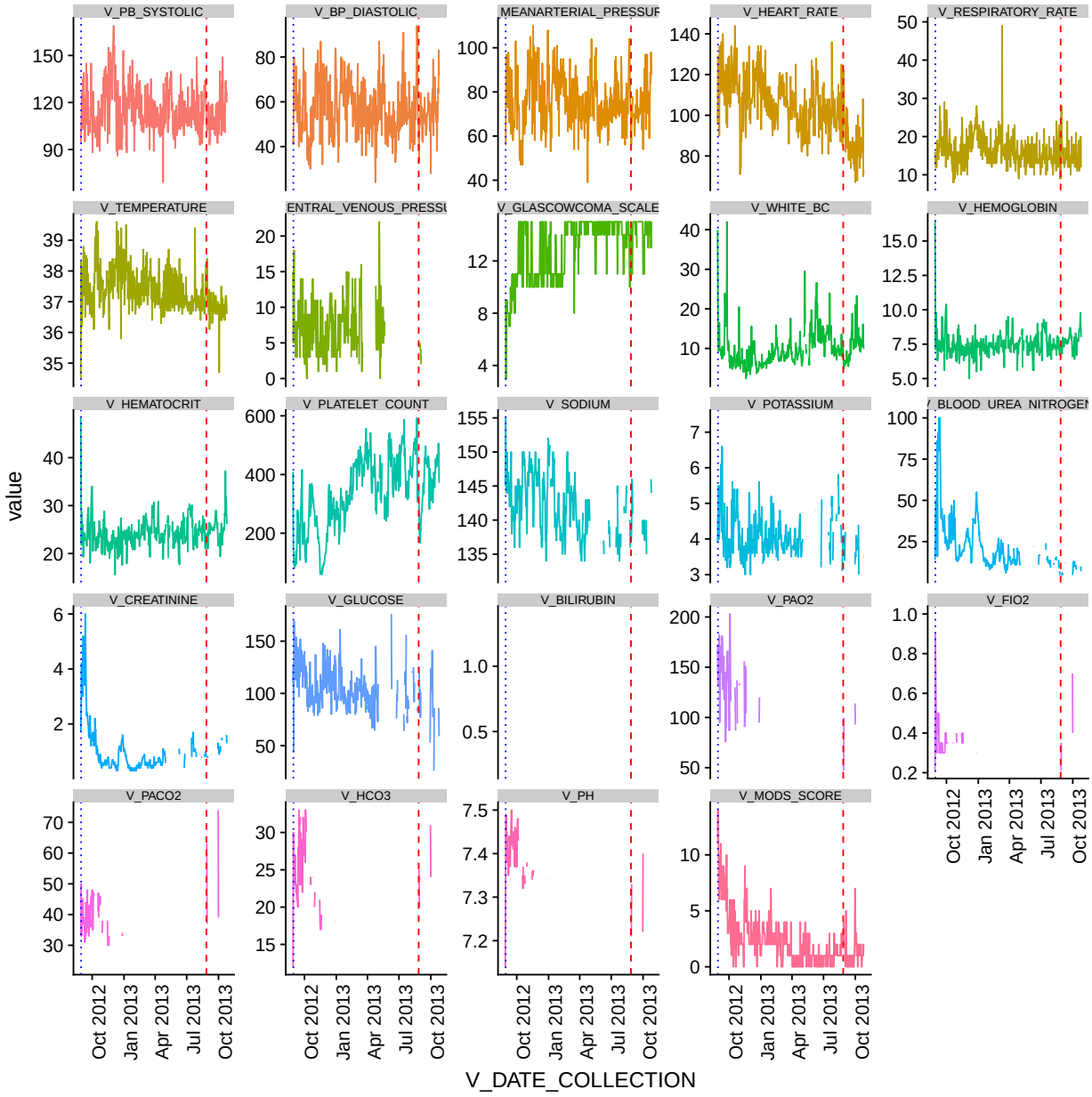
```
#Days from first collection to first infection
table(TT_per$first_any)
#Days from first collection to first blood infection
table(TT_per$first_blood)

days_from_admit_to_collection = difftime(TT_per$first_collection_date, TT_per$Admit_date, units = "days
#Days from admit to first collection
table(days_from_admit_to_collection)
#Days from admit to first infection
table(TT_per$first_any + days_from_admit_to_collection)
#Days from admit to first infection
table(TT_per$first_blood + days_from_admit_to_collection)
```

Correlation of vital signs

The charts below are for data in aggregate. Individual-level data may be more telling.

Track many patient vitals over time for an individual patient.

The red vertical line is the first blood infection and the blue is any infection.

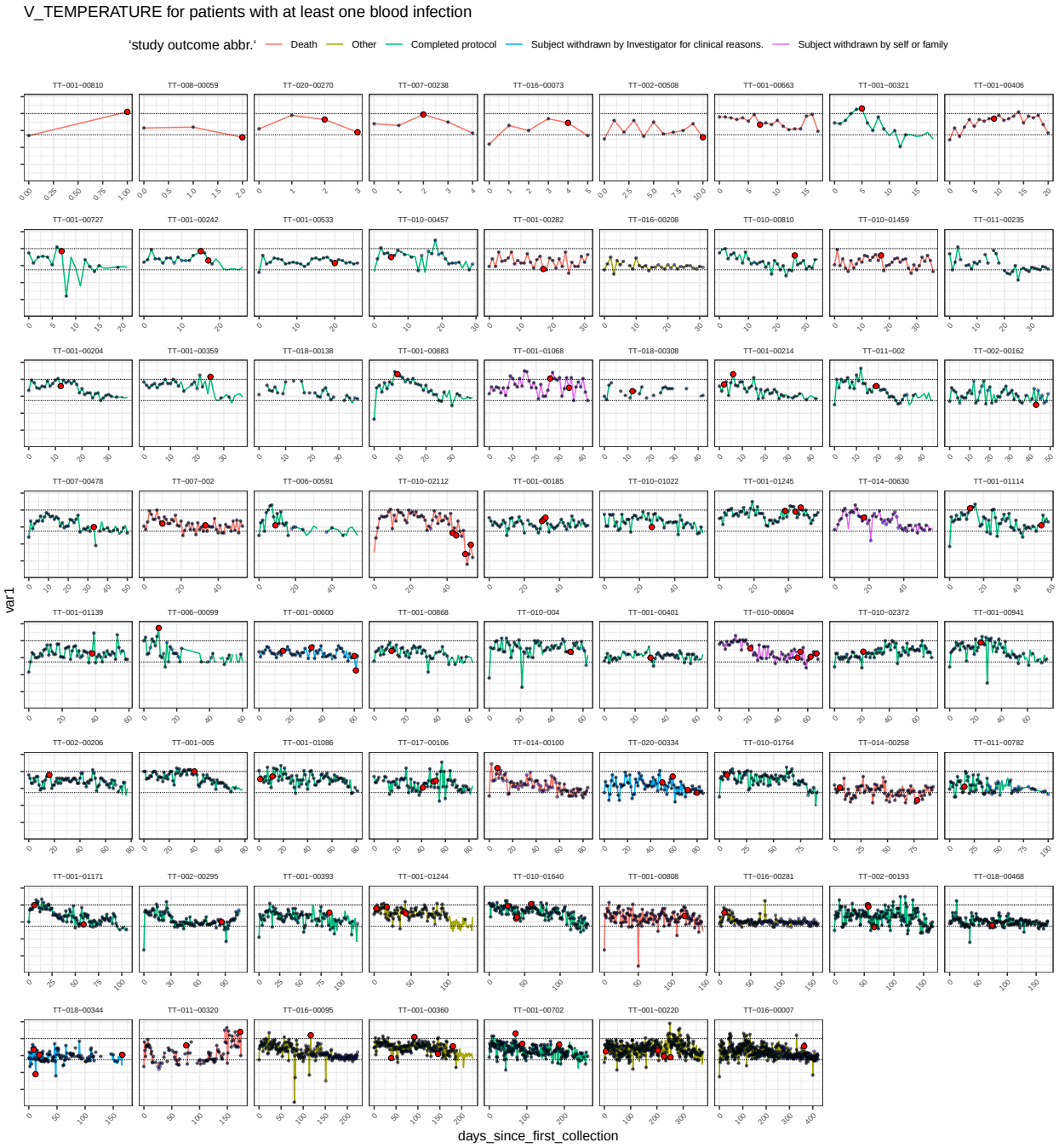


Track an individual vital over time for many patients.

The red vertical line is the first blood infection and the blue is any infection.


V_TEMPERATURE for individuals with at least one infection, all data
'study outcome abbr.' Death Other Completed protocol Subject withdrawn by Investigator for clinical reasons. Subject withdrawn by self or family
V_TEMPERATURE for individuals with at least one inf
'study outcome abbr.' Death Other
get(var1, pos = −1)
V_DATE_COLLECTION

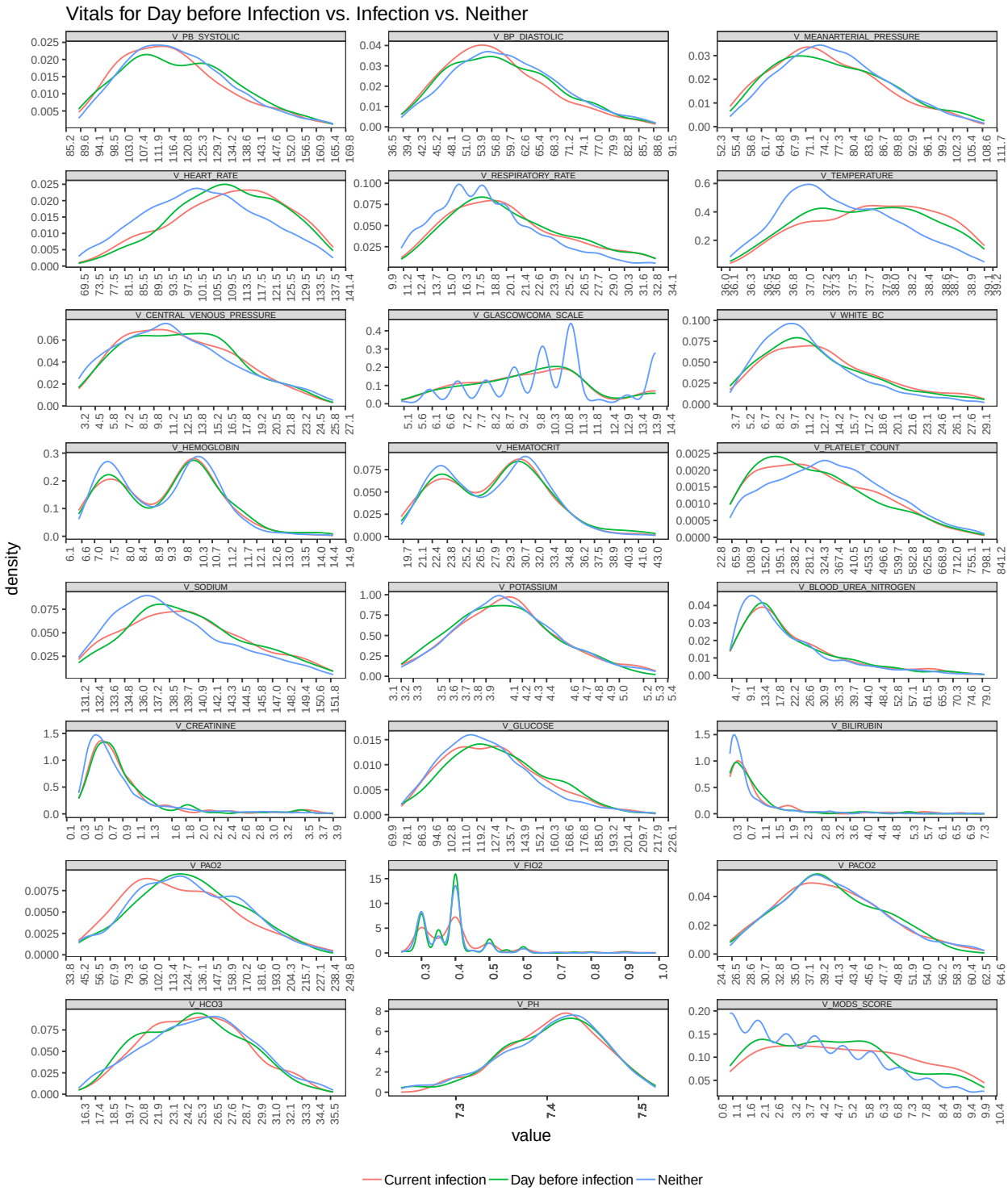
Restrict to individuals with at least one blood infection



Question 4: (pre-modeling) Which individual vital signs over time may be predictive?

Compare infection days vs. day-before infection vs. neither for individuals who have at least one infection, ***NOT*** restricted to first infection

Distributions of vital statistics by infection status excluding outliers.



Based on Figure **??** below we'd suspect that some variables may be correlated with infection or pre-infection, including heart rate, temperature, platelet count and sodium.

There are a few hundred NA entries (out of about 15,000) in "onset_tomorrow", most due to not having next-day readings for an individual, which means that the data is not missing at random. Some NA data is also due to NA entries in the onset variable.

Now RESTRICT to first infection

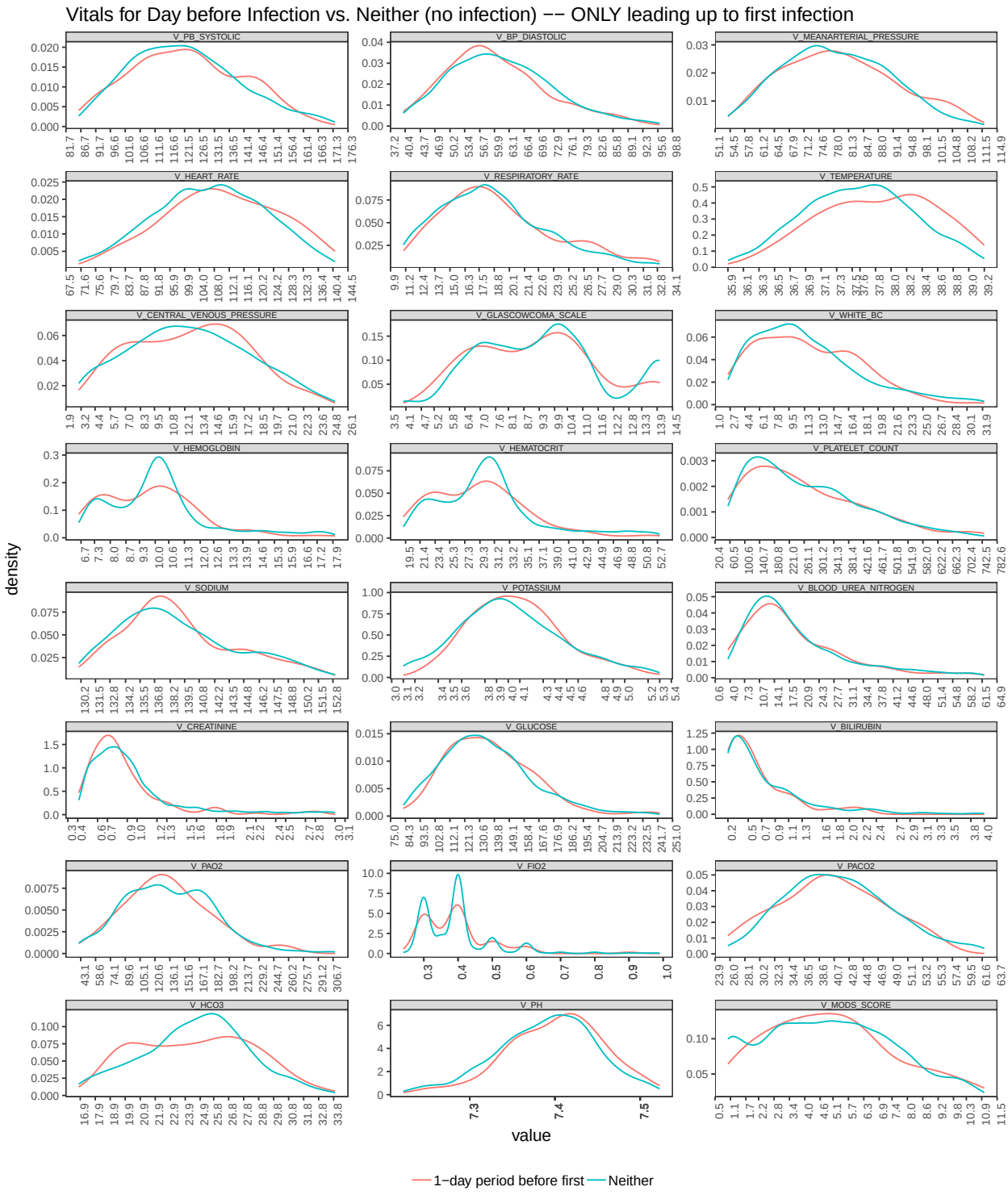
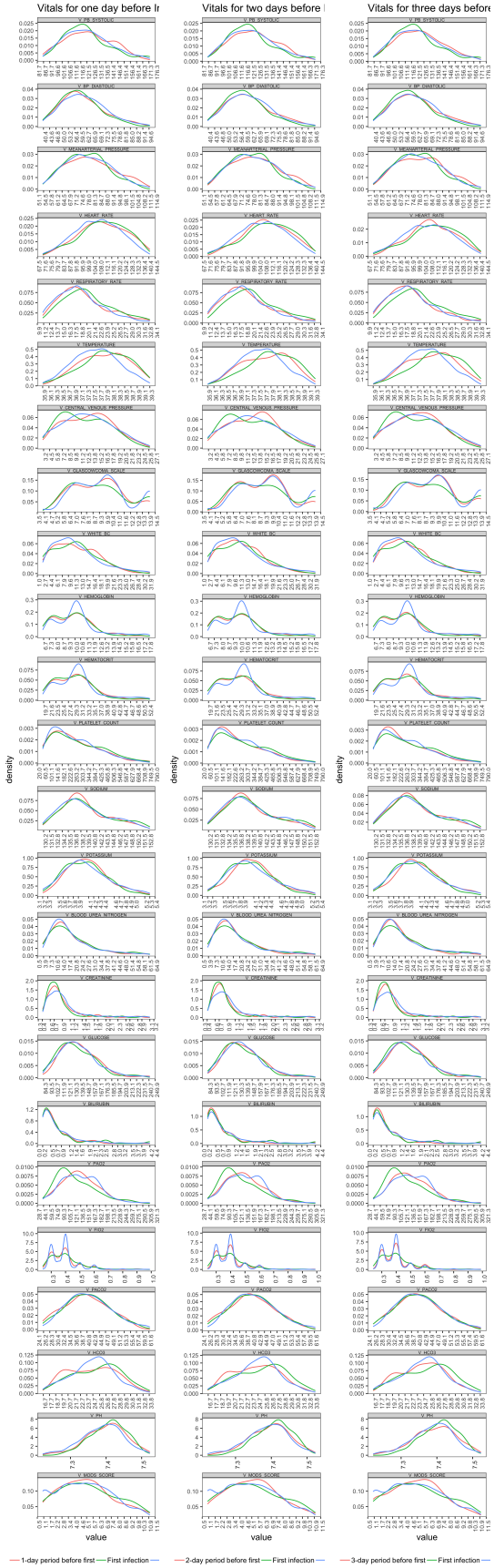


Figure 1: Distributions of vital statistics by infection status excluding outliers.

Now look at the period before infection (1, 2 or 3 day) instead of only the day before, comparing with the first infection day and non-infection days, still removing days after the first infection (note color change)



Now for BLOOD INFECTION ONLY, compare infection days vs. day-before infection vs. neither for individuals who have at least one infection, *NOT* restricted to first infection

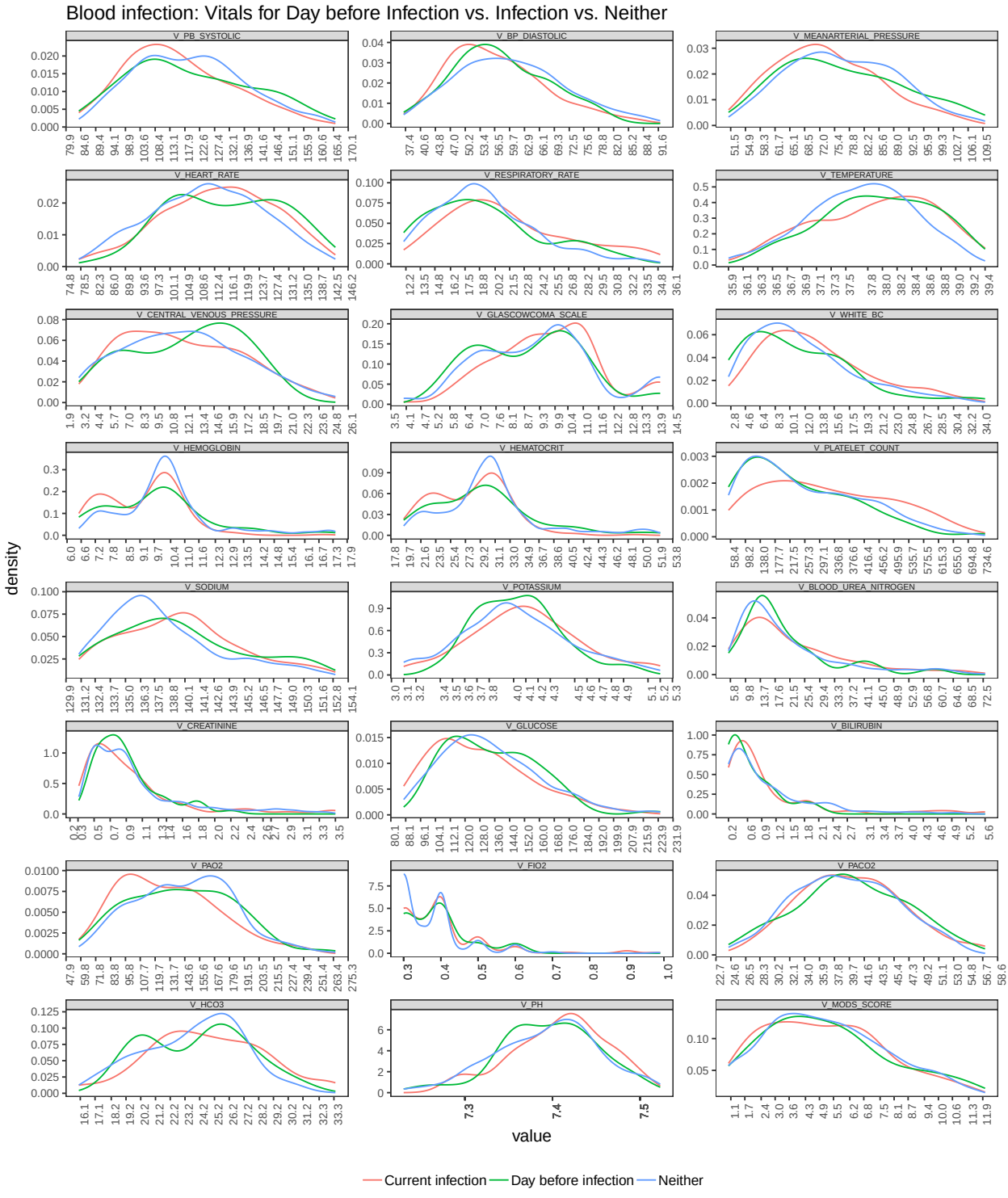
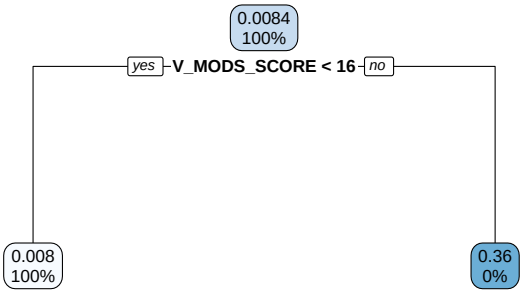


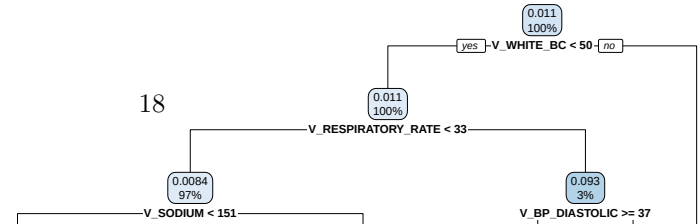
Figure 2: Distributions of vital statistics by infection status excluding outliers.

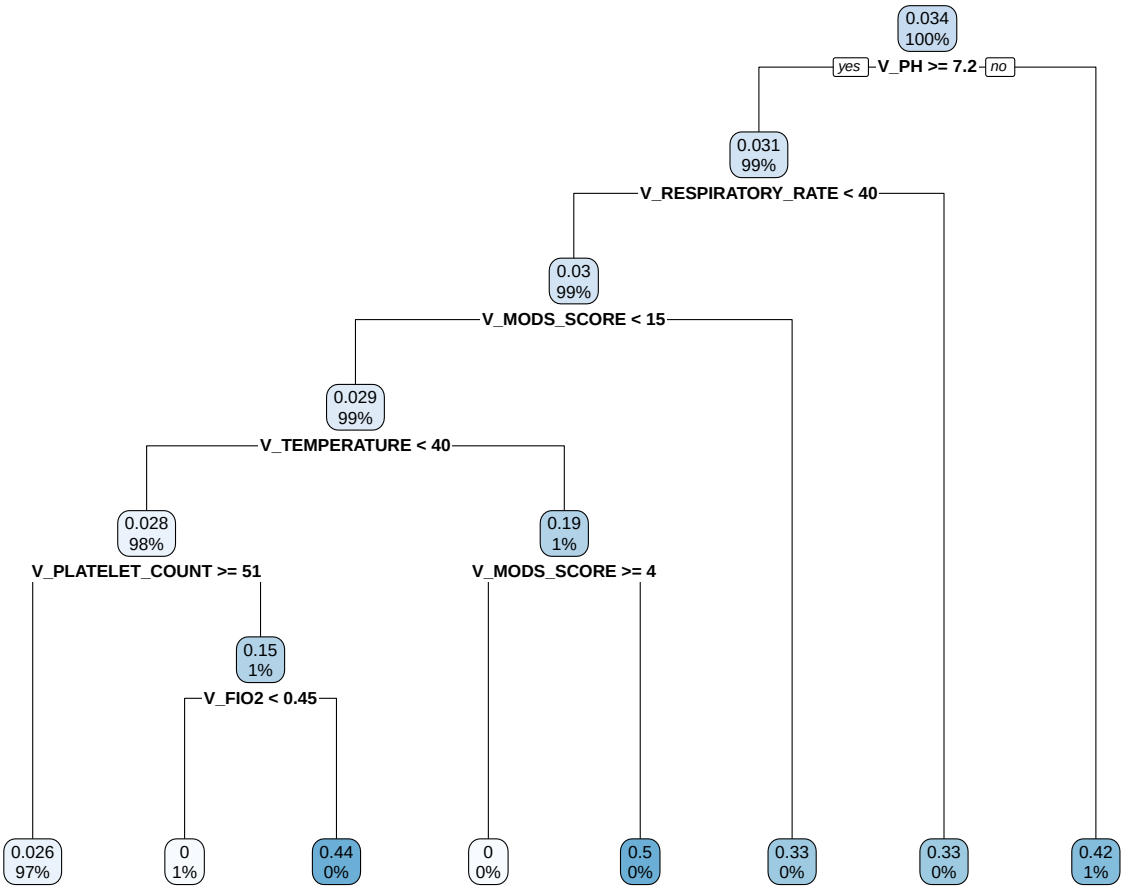
Investigation of ICU Days and PDR (predicted death rate) show that those with current or iminent infection are generally in worse health than the other groups. This is important for recognizing that factors like elevated heart rate may be due to general poor health instead of impending infection.

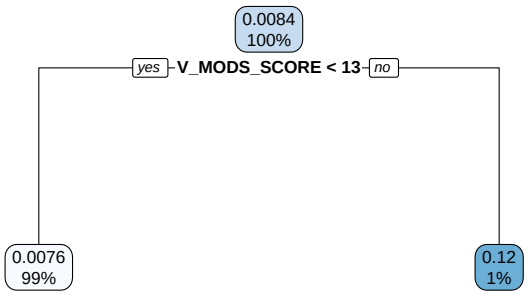
Next we look at the observed patterns more formally with a multinomial model where the possible outcomes are as labelled in Figure **??** - current infection, day before infection onset, and neither of those cases. The NA case is exluded. I used a penalized version of multinomial regression from `glmnet` in `R` with cross-validation to select variables. The tables below show the fitted non-zero coefficients using a less restrictive and more restrictive penalty term. The input data was standardized before fitting the model so that the coefficient magnitudes would be comparable, though they lose interpretability as a result.

Decision Trees






0.034
100%
yes
V_PH >= 7.2
no
0.031
99%
V_RESPIRATORY_RATE < 40
0.03
99%
V_MODS_SCORE < 15
0.029
99%
V_TEMPERATURE < 40
0.028
98%
V_PLATELET_COUNT >= 51
0.19
1%
V_MODS_SCORE >= 4
0.15
1%
V_FIO2 < 0.45
0.026
97%
0
1%
0.44
0%
0
0%
0.5
0%
0.33
0%
0.33
0%
0.42
1%



Multinomial Models:

- 1. All data.
- 2. Data to first infection and all infection days.
- 3. Data to first infection.

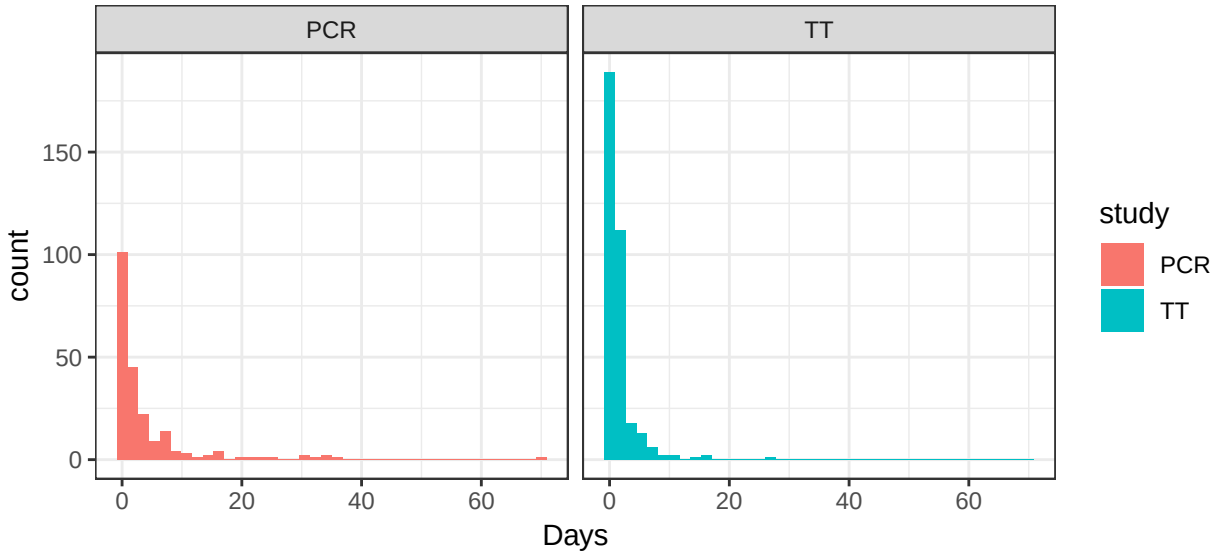
Not surprisingly, temperature has by far the largest coefficient values. Second is the MODS score (Multiple Organ Dysfunction Score), and I'm not sure what that means or if it makes sense in context. As expected by the study authors, respiratory rate, white blood cell count, and heart rate are somewhat correlated with the

outcomes. In the more limited coefficient set, only the Glascow coma scale is a negative predictor, i.e. the higher the score the less likely an infection outcome. Also, all coefficients are stronger for current infection than day before infection, with opposite sign for no infection. The same is not true of the larger set.

Another noteable outcome is that the two variables included to control for severity of condition are included in the larger model but with moderate coefficient values, and they are absent in the smaller model

PCR Study EDA

The figure compares the distribution of sepsis days per patient in the PCR study to the distribution of "onset" days per patient in the TT study. (According to the project readme, the variable "SEPSIS_STATUS" in the PCR data indicates whether the patient was determined to have a new onset of sepsis at that time point.) It's similar overall, but with more patients in the PCR study having very high numbers of sepsis days.



I will simplify this section by only comparing sepsis days to non-sepsis days, and I will see if I get a similar list of variables associated with infection as in the TT study.

Figure 3 compares the distribution of vital statistics for sepsis and non-sepsis days in patient history, after removing "outliers", i.e .the bottom and top two percent of each distribution. White blood cell count seems to behave similarly as in the TT study, but heart rate and platelet count seem to have the opposite association. For example, average heart seems to be lower for the sepsis group.

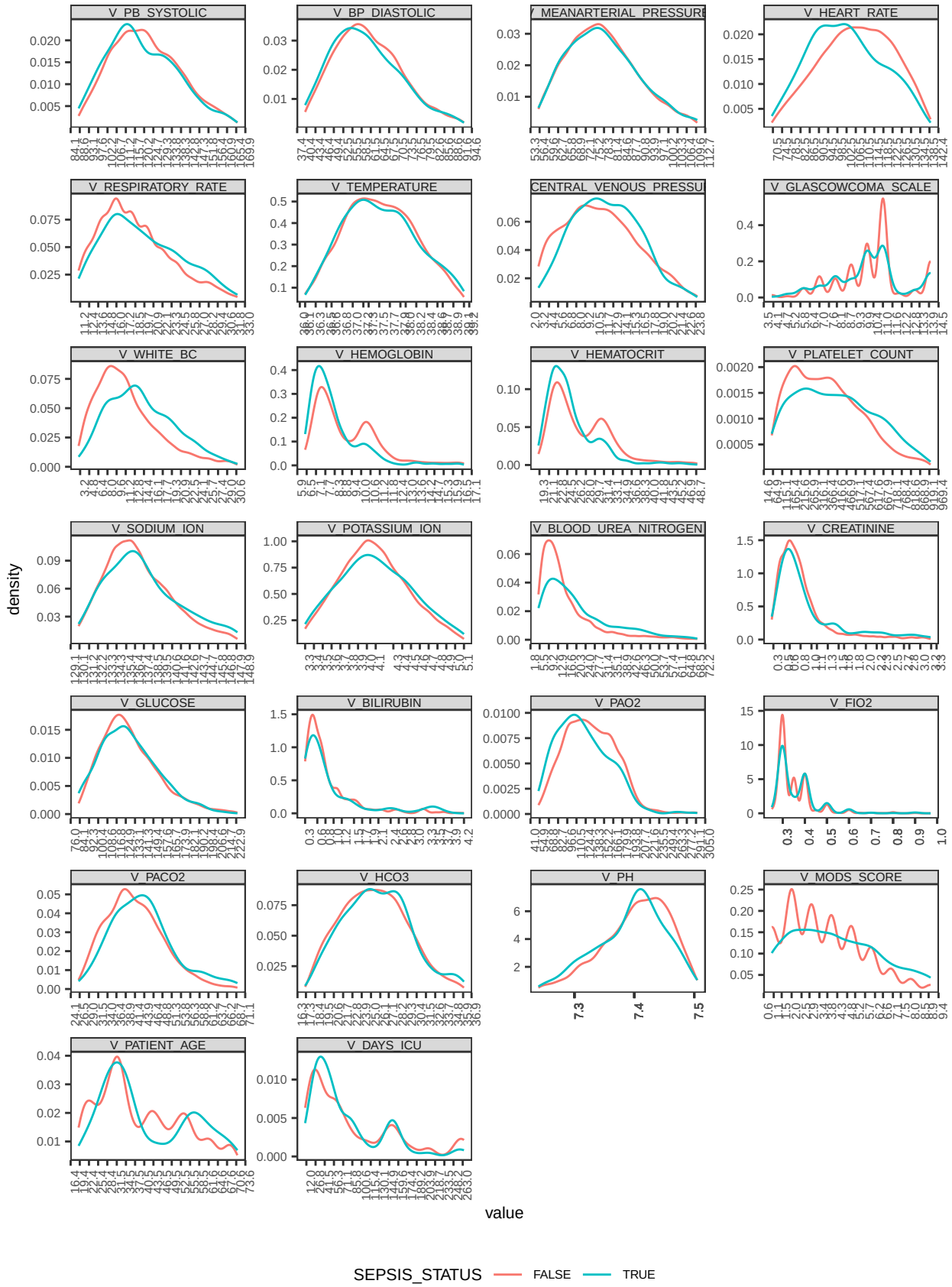


Figure 3: Distribution of vitals for sepsis vs. non-sepsis patients evaluated daily, excluding outliers

Below are fit coefficients from only the less-restrictive cross-validated multinomial model. The more restrictive model is left out since this model is already fairly small.

Note that although temperature did not appear to be significant based on the distribution plots, it is again the most important term in the model. As above, coefficients are shown on a standardized scale.

Data oddities

- A number of entries with infection Onset = "Yes" have infection Any = "No". How is that possible?

```
sum(TT$Onset=="Yes" & TT$Any=="No", na.rm = T)
```

- How should we treat the cases where infection presents in the first couple days?

```
# PCR STUDY
# TT STudy
```