A few things to note:

The dataset only consists of adults (>18 yrs old).

Create as many binary variables as possible based on the variables given (eg. is variable X normal or not?).

Questions:

1. Should sysBP and diaBP both have higher and lower bound? sysBP means systolic blood pressure. Low values indicate circulatory shock (which in turn mandates fluid resuscitation with i.v. fluids and sometimes also administration of vasopressors). High sysBP may sometimes require administration of other medications to acutely lower it. High sysBP is rarely a problem in sepsis (rather occurs in cerebrovascular insults/strokes), so I would suggest that you create a dummy at <90mmHg, where <90mmHg is basd. diaBP means diastolic blood pressure and also decreases in circulatory shock. However, diaBP is subject to individual variation (depends on stiffness of blood-vessels and more) and is more difficult to interpret. I don’t think you need to include that in your baseline model (it could possibly be worthwile to include it sometime down the road to fine-tune the model further).

2. For continuous variable such as age and weight, does it make sense to code “normal” as mean +/- 2sd? If you define normal values for age and weight in terms of your population, you would get normal values for the ICU population in the database. You can’t be certain that those values reflect the underlying source-population (i.e. general society) very well. Moreover, if you define normal like mean +/- 2sd your non-normal cases would be very extreme (i.e. constitute the 2.5% tails of the distribution). I would suggest that you create one indicator for underweight (BMI <18) and one indicator for severely obese (BMI > 35) (cutoffs according to the WHO definition).

3. The scores, like Elixhauser, SOFA and SIRS, does it make sense to make them binary, like very severe vs. not so severe?

I think that there is a risk that you miss out on quite a bit of important information if you just dichotomize the scores. Moreover, the SIRS score is affected by being quite non-specific (i.e. a lot of people with diseases that are much less severe than sepsis also display one or more SIRS criteria). I have outlined a strategy that might work in the table below.

4. Things in red are the ones still having trouble to binarize.

|  |  |  |
| --- | --- | --- |
| **Label** | **Meaning** | **Normal range (if applicable)** |
| **OBSERVATIONS** | | |
| Gender | -- | binary |
| Age | -- | I’d suggest to use indicator variables for age 18-65, 65-85 and >85. From my experience, it’s most important to identify the really old (who have a lot of co-morbidities and scarce reserves), which you do in the group >85. |
| Elixhauser | Elixhauser Comorbidity Index(0-14) | Continuous in dataset  I’d suggest that you use score ≥3 as an indicator for severe comorbidity (in the original paper where the coding algorithm was developed, 17.8% of observations had score ≥3 and 11.7% of these died in the hospital) |
| Re\_admission | -- | binary |
| SOFA | sequential organ failure assessment score (0-23) | If you can’t use one-hot encoding I suggest… |
| SIRS | Systemic inflammatory response syndrome (0-4) | SIRS is a bit tricky. Sepsis is defined as having SIRS ≥1 while being infected by a microorganism. If you are interested in sepsis-patients only, you should therefore probably exclude all patients with SIRS = 0. The thing is that many people exhibit one or more SIRS criteria due to other diseases, that are not always very severe. Similarly, patients who fulfil 2 or 3 SIRS criteria do not necessarily need to be much sicker than patients who fulfil 1. Therefore, I think that it might not be a very good strategy to use SIRS scores only, to grade the severity of sepsis. Instead, I would suggest that you create a dummy that indicates septic shock. Septic shock is defined as SIRS ≥ 1 AND sysBP <90mmHg that persists despite intravenous fluid therapy OR arterial\_lactate >1.0. One way of coding septic shock that takes into account the ‘persists’ criterion could be ‘SIRS ≥1 AND (sysBP <90mmHg AND received intravenous fluids during the previous 4h time window) OR arterial\_lactate>1.0’. If you want to capture even more granularity, you could create a dummy-variable indicating severe sepsis (which is more severe than sepsis only, but less severe than septic shock), to end up with a total of 3 categories. Severe sepsis can be coded as ‘SIRS≥1 AND signs of organ dysfunction’. I have pasted a table below, that indicates criteria for organ dysfunction. You should be able to construct most of them using this database. Note that the variables I outlined in this section may display collinearity with ‘Shock index’ variable below. |
| Weight\_kg | -- | I’d suggest that you convert to BMI and create one indicator variable for BMI<18 and one for BMI≥35. |
| GCS | Glasgow Coma Scale: correlates with the severity of brain injury and prognosis | Maximum score is 15 which has the best prognosis  Minimum score is 3 which has the worst prognosis  **Scores of 8 or above** have a good chance for recovery  GCS is an indicator of mental status. If you are below 8 you are in a coma. Hence, if you have sepsis and GCS of 8, you are really sick. I would suggest that you identify GCS ≥13 as positive (these patients are quite alert). <8 could be used as a dummy for negative prognosis (and =3 as a dummy for super-bad prognosis). There is one potential difficulty with regards to GCS. The thing is that I think that doctors assess GCS also in patients who are sedated (i.e. put to sleep with strong medications). Most patients who have mechanical ventilation will be sedated. Hence the variable can be difficult to interpret in patients who are on mechanical ventilation. |
| HR | Heart rate | 60 to 100 beats a minute  Often times, tachycardia (increased heart rate, which might indicate circulatory shock) is defined as >90 beats per minute. If you choose the cutoff at >100 you will identify a slightly sicker subset. The advantage with using >90 is that it conforms to the SIRS criteria that are widely used in sepsis-research. If one wants to make the variable even fancier, one could look into normal levels for the patient’s age group. |
| SysBP |  | Below 120 I’d suggest that you create a dummy indicating ≤90mmHg (which indicates circulatory shock and is bad for the patient). Is this one timestamp for each 4h period btw? If not, it might be tricky to code if there are much fluctuations within each 4h interval. |
| MeanBP | Some kind of average? centered at 75, majority between 50-100. | This may be very collinear with SysBP and I’d suggest you don’t use this in the first version of the model. |
| DiaBP |  | Below 80 Since the definition of circulatory shock is foremost based on SysBP, I’d suggest you don’t use this in the first version of the model, but go with SysBP instead. |
| Shock\_Index | Between 0-2 continuous. | I am not sure which index this refers to. If it is a validated index addressing septic shock, it could carry quite some information and be worthwile to include. One should then be aware of the risk of it displaying collinearity with the septic shock variable I suggested in conjunction to the ‘SIRS’ discussion above. |
| RR | Respiratory rate | 12 to 20 breaths per minute. >20/minute indicates tachypnea (i.e. ‘fast respiratory rate’) which is indicative of bad prognosis. |
| SpO2 | peripheral oxygen saturation | 95% – 100%. While ≥95% is normal, my experience is that some elderly people and some others sometimes are <95% without being very sick. <90% is often indicating disease. However, patients with COPD may sometimes be lower than that in their steady state. Since you have the ‘paO2’ variable below, I’d suggest that you use that instead of SpO2 (‘paO2’ also indicates the oxygen level in the blood, but is measured directly in the blood and is therefore much more exact). You risk collinearity if you use both SpO2 and paO2. |
| Temp\_C | Degree Celsius | 96.8 °F (36°C) to 100.4°F (38°C) I’d suggest you create an indicator for <36 and one for >38. Bear in mind that this is one of the SIRS criteria, so if you include dummies for each SIRS-score or SIRS as a continuous variable you may experience collinearity. This should be less of a problem if you use the ‘septic shock’ indicator I suggested instead of the SIRS criteria to indicate severity of disease. |
| FiO2\_1 | Related to mechanical ventilator (physician’s choice) | I think that this is not super-important. What is instead important is the ratio between FiO2 and PaO2 (the ratio says how much of the oxygen that the doctor puts into the patient that actually enters the bloodstream of the patient). You have that variable in ‘PaO2\_FiO2’ below. |
| Potassium | -- | 3.5-5.0 mEq/L |
| Sodium | -- | 135 to 145 mEq/L You could consider to create a dummy indicating hyponatremia (low sodium). The normal values depend somewhat on age and gender, but I’d suggest you use Sodium ≤120 to capture the severe cases. |
| Chloride | -- | 96 to 106 mEq/L |
| Glucose | -- | The American Diabetes Association recommends a fasting plasma glucose level of 70–130 mg/dL (3.9-7.2 mmol/L) and after meals less than 180 mg/dL (10 mmol/L). |
| BUN | Blood Urea Nitrogen (also known as Urea or Urea nitrogen) | 7 to 20 mg/dL (2.5 to 7.1 mmol/L). This is used a lot in the US, but not as much in Sweden. I am not entirely familiar with which values would indicate very severe disease. |
| Creatinine | -- | 0.6 to 1.2 milligrams (mg) per deciliter (dL) in adult males and 0.5 to 1.1 milligrams per deciliter in adult females. Creatinine is a common indicator for kidney-failure (which if it happens, mandates RRT, as we spoke about yesterday). The thing is that creatinine depends on a lot of things (body composition fat/muscle, age and in the ICU setting importantly also the patient’s fluid status (i.e. creatinine can be artificially low if the patient has received a lot of intravenous fluids). If you want to experiment with creatinine sometime, I’d suggest that you try to find the baseline values of the patients (which can be tricky to obtain due to data-privacy laws etc). One could consider modelling increases between 4h time-windows, but that may be confounded by RRT. |
| Magnesium |  | 1.5-2.5 mEq/L I think this variable may not have the highest information yield. |
| Calcium |  | 8.5-10.2 mg/dL Very high values could identify patients with hypercalcemic crises, which are quite severe. The scale for measurement is different from that in Sweden though, so I am not sure exactly where to put the cutoff (but maybe you can google ‘hypercalcemic crisis’?). The variable will likely be collinear with Ionised\_Ca below and I’d suggest you only use one of them. |
| Ionised\_Ca |  | 4.64 to 5.28 mg/dL Very high values could identify patients with hypercalcemic crises, which are quite severe. The scale for measurement is different from that in Sweden though, so I am not sure exactly where to put the cutoff. |
| CO2\_mEqL | Blood CO2 | 23 to 29 mEq/L This coud be important as high values indicate CO2 retention (which can indicate severe breathing-problems). However, I’m not familiar with the scale. |
| SGOT | Aspartate aminotransferase, an enzyme made by liver, leaks out to blood stream during liver damage | 5 to 40 units per liter of serum This can be tricky to interpret as some chronic cases of liver failure (which have very poor liver-function) may still have quite low values (they haven’t got many more liver-cells that can release the enzyme). In a similar fashion, some acute toxicities with high values may still have quite a bit of liver-function left. This also applies to SGPT below. |
| SGPT | alanine aminotransferase, made by liver | 7 to 56 units per liter of serum |
| Total\_bili | Total bilirubin | 0.1 to 1.2 mg/dL (1.71 to 20.5 µmol/L) Can be a good indicator of various hepatic diseases, but not necessarily a good indicator for mortality. |
| Albumin | Albumin | 3.5 to 5.5 g/dL or 35-55 g/liter This could be a good indicator of liver-function in some circumstances. However, it is confounded by malnutrition and some other conditions. If you want to capture liver-related disease/mortality I’d instead suggest that you instead construct a MELD score (which would be quite impressive and unique, and totally achievable given the richness of the database). I think MELD has showed AUC >0.85 for in-hospital death in patients with liver-failure in some populations. MELD is calculated as follows: 9.57\*ln(creatinine mg/dL)+3.78\*ln(total bilirubin mg/dL)+11.2\*ln(INR)+6.43. There is also a version taking sodium into account, but I believe the original may be the most frequently used. |
| Hb | Hemoglobin | For men, 13.5 to 17.5grams per deciliter. For women, 12.0 to 15.5 grams per deciliter. Hb is difficult to interpret, as it is subject to the same dilution-issues as creatinine. However, Values <7.0 should generally indicate that something is quite wrong. |
| WBC\_count | White blood cell count | 4,500 to 11,000 WBC per microliter (4.5 to 11.0 × 10^9/L) This is tricky to interpret, as both increases and decreases can indicate disease. Some diseases that are less severe can still have higher WBC count than other more severe diseases. This is also a SIRS criterion, which is why I’d suggest that you create the ‘septic shock’ variable I outlined above and leave WBC count out of the model. |
| Platelets\_count |  | 150,000 to 450,000 Low values carry a high risk of bleeding. I’d suggest that you construct a dummy indicating platelets <30. |
| PTT | Partial Thromboplastin Time (how long it takes the blood sample to clot after adding the chemicals) | 25 to 35 seconds This can be used in defining ‘coagulation abnormalities’ in conjunction to ‘severe sepsis’ (outlined above). See closer description in pasted table below. |
| PT | Prothrombin time | 11 to 13.5 seconds Use PTT to define ‘coagulation abnormalities’ instead. |
| INR | international normalized ratio | INR of 1.1 or below is considered normal This can be used in defining ‘coagulation abnormalities’ in conjunction to ‘severe sepsis’ (outlined above). See closer description in pasted table below. |
| Arterial\_pH |  | 7.35 to 7.45 Low values will likely be a good indicator of disease. There is some individual variation, but I’d suggest you create at least two dummies, one indicating pH <7.00 and one indicating pH [7.00-7.15). It is a bit tricky to outline exactly where to put the cutoff against normal, since patients with pH 7.15 are quite sick, but patients with pH 7.30-7.35 can be quite ok (that depends on what is the underlying cause of low pH). One suggestion (without having spent much time looking for references) is [7.15-7.30). As pH increase is a physiological effect of hyper-ventilation, you will likely see some high pH values in the database. This can happen consciously in the ICU and I think you should avoid including a dummy for high pH. |
| paO2 | partial pressure of arterial oxygen | * greater than 80 mmHg/10.6 kPa * 90 - 95 mm Hg.   >10.6kPa would be normal for most people. <8.0 (I have no exact reference for this) would indicate that you have an oxygen-problem of some dignity. |
| paCO2 | partial pressure of arterial carbon dioxide | 35 - 45 mmHg (4.7 - 6.0 kPa) When it comes to paCO2, this is most often a problem if it is high (low values indicate hyper-ventilation, which could be due to a variety of reasons). >6.0 would likely capture difficulties in getting rid of CO2 in most people (higher values are worse, but I can’t come up with an exact cutoff. What one could expect is also quite dependent on if the patient has COPD. In patients without COPD, >6.0 would definitely indicate disease). |
| Arterial\_BE | Base excess: is derived from pH, Hb, body temperature, and pCO2. BE quantifies the metabolic portion and essentially tells us how much acid or base is needed to titrate patient to pH=7.4 if the pCO2 is held steady at 40 mmHg. | -4 to +4 mmol/L This is less important when you already include pH, paCO2 and paO2. |
| Arterial\_lactate |  | 0.5-1 mmol/L. Very sick patient’s normal: <2.0 mmol/L This is sometimes used as a prognostic marker for death in sepsis. If you use it to construct the ‘severe sepsis’ variable I suggested, definitions say >1.0 mmol/L. Since it is a continuous measure and a lot of ICU patients are sick and have elevated values, I think it could make sense to create a dummy-variable indicating >2.0mmol/L also (but not to be used in defining ‘severe sepsis’). |
| HCO3 |  | 22 to 28 mEq/L This is less important when you already include pH, paCO2 and paO2. |
| PaO2\_FiO2 |  | Approx. 500 <300 indicates organ-dysfunction of the lung according to the ‘severe sepsis’ criteria posted in the table below. Beware of potential collinearity if you use a dummy indicating this while also using a ‘severe sepsis’ variable. |
| output\_total | Urine output (majority<15,000) | I wonder if this could be urine output per 24h or something like that. I’d suggest going with the output\_4hourly variable below, since its relation to time is much more clear. |
| output\_4hourly | Urine output every 4 hours (majority<1,000) | 800 to 2000 milliliters **per day** (with a normal fluid intake of about 2 liters per day) I’d suggest you use this to define acute oliguria (i.e. urine output <0.5ml/kg/hour). I think you would see that happen before people are put on RRT. |
| cumulated\_balance\_tev | ??majority between -10,000~+20,000 | Sounds like cumulative urine output. Sounds like it could refer to cumulative urine output during the ICU stay, which would be less useful since it is then also confounded by time. |
| **INTERVENTIONS** | | |
| median\_dose\_vaso |  | continuous |
| max\_dose\_vaso |  | continuous |
| input\_total\_tev |  | continuous |
| input\_4hourly\_tev |  | continuous |
| sedation |  | binary |
| mechvent |  | binary |
| rrt | renal replacement therapy (dialysis) | binary |
| **OUTCOME** | | |
| died\_in\_hosp |  | Binary (patient dies in hospital, but not necessarily in ICU) |
| mortality\_90d |  | Binary (likely refers to all cause mortality within 90 days, i.e. also mortality that happens after the patient is discharged from the hospital). |

mechvent: once connect, usually lasts for a while, like at least a day, or more. Do not change back forth very much.

rrt: lasts a few hours or longer, but may have to do the next day. So more likely to see on-and-off switch.

|  |  |
| --- | --- |
| **Organ-dysfunction variables table** |  |
| Arterial hypoxemia (ratio of the partial pressure of arterial oxygen to the fraction of inspired oxygen, <300) | Can be calculated using ‘PaO2\_FiO2’ variable |
| Acute oliguria (urine output, <0.5 ml/kg/hr or 45 ml/hr for at least 2 hr) | Can be calculated using ‘output\_4hourly’ variable |
| Increase in creatinine level of >0.5 mg/dl (>44 μmol/liter) | This must be calculated with reference to the patient’s baseline level (which might be tricky to find) |
| Coagulation abnormalities (international normalized ratio, >1.5; or activated partial-thromboplastin time, >60 sec) | Can be calculated using ‘INR’ variable or ‘PT’ variable |
| Paralytic ileus (absence of bowel sounds) | This appears not registered in the database |
| Thrombocytopenia (platelet count, <100,000/mm3) | This can be calculated using the ‘Platelets\_count’ variable |