Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

ricu: R's interface for ICU data

Nicolas Bennett
ETH Zürich

Drago PLecko
ETH Zürich

Ida-Fong Ukor
East Kent Hospitals

Abstract

The abstract of the article.

Keywords: medicine, intensive care, computational bioinformatics.

1. Introduction

Collection of electronic health records has seen a significant rise in the recent years Evans (2016), opening up an opportunities for a large body of data-driven research oriented towards improving patient care and outcomes, together with helping clinicians in decision-making Jiang, Jiang, Zhi, Dong, Li, Ma, Wang, Dong, Shen, and Wang (2017).

One example of a problem that has received much attention from the machine learning community is early prediction of sepsis in ICU Desautels, Calvert, Hoffman, Jay, Kerem, Shieh, Shimabukuro, Chettipally, Feldman, Barton *et al.* (2016); Nemati, Holder, Razmi, Stanley, Clifford, and Buchman (2018); Futoma, Hariharan, Sendak, Brajer, Clement, Bedoya, O'Brien, and Heller (2017); Kam and Kim (2017). Interestingly, there is evidence that a large proportion of the publications are based on the same dataset Fleuren, Klausch, Zwager, Schoonmade, Guo, Roggeveen, Swart, Girbes, Thoral, Ercole *et al.* (2019), the Medical Information Mart for Intensive Care (MIMIC) Johnson, Pollard, Shen, Li-wei, Feng, Ghassemi, Moody, Szolovits, Celi, and Mark (2016), which shows a systematic lack of external validation. Part of this problem might well be the lack of a computational infrastructure handling multiple datasets. The MIMIC-III dataset consists of 26 different tables containing about 20GB of data. While much work and care has gone into data preprocessing in order to provide a self-contained ready to use data resource with MIMIC, seemingly simple tasks such as computing a SOFA score for patients (Vincent, Moreno, Takala, Willatts, De Mendonça, Bruining, Reinhart, Suter, and Thijs 1996), remains a non trivial effort. This is only exacerbated when aiming to co-integrate multiple different datasets of this form is, spanning hospitals and even countries in order to capture effects of differing practice and demographics.

The aim of the **ricu** package is to provide computational infrastructure allowing users to access complex research questions as easily as possible. The package also aims to enable users to write dataset-agnostic code which can simplify implementation and shorten the necessary time for prototyping code to different datasets. In particular, the package handles three large, publicly available intensive care databases: the already mentioned MIMIC-III database from the Beth Israel Deaconess Medical Center in Boston, Massachusetts, the eICU Collaborative Research Database Pollard, Johnson, Raffa, Celi, Mark, and Badawi (2018), containing data collected from 208 hospitals across the United States, and the HiRID database Faltys (2018) from the Department of Intensive Care Medicine of the Bern University Hospital, Switzerland. Together with this, much of the functionality used is also aimed to accommodate for addition of possible additional datasets, provided by the user. The work most similar to ours is that of Adibuzzaman, Musselman, Johnson, Brown, Pitluk, and Grama (2016) and Wang, McDermott, Chauhan, Ghassemi, Hughes, and Naumann (2020). However, these works address only the MIMIC-III dataset and do not have an emphasis on dataset inter-operability.

The structure of the manuscript is as follows. In Section 2 we outline the different types of data useful for research related to intensive care medicine. We explain the most important parts of the package functionality which are used to handle the different data types. In Section 3 we provide simple examples which illustrate how some simple research questions can be explored in only a couple of lines of code.

2. Implementation

In this Section we go over the categories of data useful for research problems related to intensive care medicine. The categories we define are fairly broad and somewhat loosely defined, as this is not the main focus of the manuscript.

2.1. Physiological data

Labs, vitals. Could introduce the `ts_tbl` here.

2.2. Treatment-related information

Antibiotics, vasopressors, mechanical ventilation... could introduce the `win_tbl` here.

2.3. Co-morbidities

Based on ICD-9 codes. Should enable the extraction of co-morbidities used for the Charlson and Elixhauser scores.

2.4. Admission diagnoses

Categorizing into surgical, non-surgical and other might be sufficient for now.

2.5. Patient information

Age, gender, other demograpichs, patient stay information.

2.6. Outcomes

Death outcome, prolonged ICU stay outcome.

3. Examples

We focus on two simple examples with which we try to cover most of the data types described in Section 2.

3.1. Lactate and mortality

The first example we look at is the association of lactate levels and mortality. This problem has been studied before and it is widely accepted that both static and dynamic lactate indices are associated with increased mortality (Haas, Lange, Saugel, Petzoldt, Fuhrmann, Metschke, and Kluge 2016; Nichol, Bailey, Egi, Pettila, French, Stachowski, Reade, Cooper, and Bellomo 2011; Van Beest, Brander, Jansen, Rommes, Kuiper, and Spronk 2013). We quickly look at how one might fit a time-varying proportional hazards Cox model (Therneau and Lumley 2015) in order to investigate this problem. We additionally include the Sequential Organ Failure Assessment (SOFA) score (Vincent *et al.* 1996) as a general predictor of illness severity.

```
R> source <- "mimic_demo"
R> # data loading
R> tbl <- load_concepts(c("lact", "death", "sofa"), source,
+                       verbose = FALSE)

Only concept data from a single data source can be loaded a the time. Please
choose one of mimic_demo

Data aggregation cannot be disabled (i.e. passing an ‘aggregate‘ value of
‘FALSE‘ for at least one concept) when data merging is enabled.

No overlap between configured id var options and available columns for table
‘labevents‘ in mimic_demo

cannot mix interval lengths when row-binding

No overlap between configured id var options and available columns for table
‘admissions‘ in mimic_demo

cannot mix interval lengths when row-binding

No overlap between configured id var options and available columns for table
‘labevents‘ in mimic_demo

cannot mix interval lengths when row-binding
```

```
removed 32 (2.77%) of rows due to out of range entries

not all units are in [mmHg]: mm Hg (99.82%), MM HG (0.18%)

No overlap between configured id var options and available columns for table
‘labevents‘ in mimic_demo

cannot mix interval lengths when row-binding

removed 4 (1.18%) of rows due to out of range entries

not all units are in [%]: NA (18.26%)

cannot mix interval lengths when row-binding
cannot mix interval lengths when row-binding
cannot mix interval lengths when row-binding

x does not contain column ‘starttime‘

No overlap between configured id var options and available columns for table
‘labevents‘ in mimic_demo

cannot mix interval lengths when row-binding

removed 3 (0.19%) of rows due to ‘NA‘ values

No overlap between configured id var options and available columns for table
‘labevents‘ in mimic_demo

cannot mix interval lengths when row-binding
cannot mix interval lengths when row-binding

removed 22 (0.14%) of rows due to ‘NA‘ values

removed 13 (0.08%) of rows due to out of range entries

cannot mix interval lengths when row-binding

removed 771 (33.52%) of rows due to ‘NA‘ values

removed 1 (0.07%) of rows due to out of range entries

cannot mix interval lengths when row-binding

removed 2 (50%) of rows due to ‘NA‘ values
```

```
cannot mix interval lengths when row-binding

removed 371 (36.16%) of rows due to ‘NA‘ values

cannot mix interval lengths when row-binding

removed 9 (27.27%) of rows due to ‘NA‘ values

cannot mix interval lengths when row-binding

removed 2 (0.05%) of rows due to ‘NA‘ values

cannot mix interval lengths when row-binding

removed 5 (0.13%) of rows due to ‘NA‘ values

cannot mix interval lengths when row-binding

removed 7 (0.18%) of rows due to ‘NA‘ values

cannot mix interval lengths when row-binding

removed 2 (0.1%) of rows due to ‘NA‘ values

cannot mix interval lengths when row-binding

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

cannot mix interval lengths when row-binding

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

cannot mix interval lengths when row-binding

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘
```

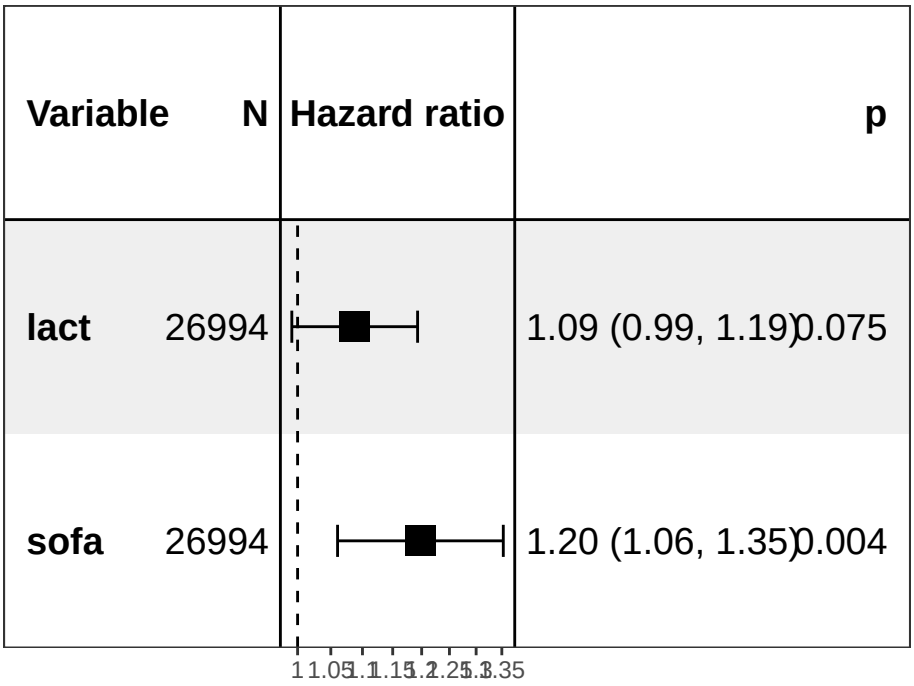
```
x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

No overlap between configured id var options and available columns for table
‘labevents‘ in mimic_demo

cannot mix interval lengths when row-binding

removed 1 (0.06%) of rows due to ‘NA‘ values

removed 1 (0.06%) of rows due to out of range entries

cannot mix interval lengths when row-binding

removed 20 (0.22%) of rows due to ‘NA‘ values

removed 1 (0.01%) of rows due to out of range entries

not all units are in [mL]: NA (0.08%)

x does not contain column ‘charttime‘
x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

cannot mix interval lengths when row-binding

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘
```

```
x does not contain column ‘icustay_id‘

cannot mix interval lengths when row-binding

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

cannot mix interval lengths when row-binding

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

cannot mix interval lengths when row-binding

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

cannot mix interval lengths when row-binding

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

cannot mix interval lengths when row-binding

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

cannot mix interval lengths when row-binding

x does not contain column ‘icustay_id‘
```

```
x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

cannot mix interval lengths when row-binding

x does not contain column ‘charttime‘
x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

cannot mix interval lengths when row-binding

x does not contain column ‘icustay_id‘

x does not contain column ‘charttime‘

x does not contain column ‘icustay_id‘

cannot mix interval lengths when row-binding
```

```
R> tbl <- tbl[, c(meta_vars(tbl), "lact", "sofa", "death"), with = FALSE]
R> tbl <- tbl[, lact := nafill(lact, "locf")]
R> tbl <- tbl[, lact := nafill(lact, fill = 1)]
R> tbl[, event := as.integer(sum(death, na.rm = TRUE) > 0), by = c(id_vars(tbl))]
R> tbl[, event := last_event(event), by = c(id_vars(tbl))]
R> tbl[, next_charttime := charttime+1L]
R> # model fitting
R> cox_time_mod <- coxph(
+   Surv(charttime, next_charttime, event) ~ lact + sofa,
+   data = tbl
+ )
```

We visualize the results of the model



A simple exploration already shows that the increased values of lactate are associated with mortality, even after adjusting for the SOFA score.

3.2. Diabetes and insulin treatment

The next example we turn to covers the usage of co-morbidities and treatment related information. We look at the amount of insulin administered to patients in the first 24 hours from their ICU admission. In particular, we investigate if patients who are diabetic receive more insulin in the first day of their stay. We extract the data as follows:

```
R> source <- "mimic_demo"
R> ins_breaks <- c(0.01, 10, 20, 30, 40)
R>
R> cohort <- stay_windows(source)
R> ins_treat <- load_concepts("ins", source, verbose = FALSE)

Only concept data from a single data source can be loaded a the time. Please
choose one of mimic_demo

Data aggregation cannot be disabled (i.e. passing an ‘aggregate‘ value of
‘FALSE‘ for at least one concept) when data merging is enabled.

x is not a ‘difftime‘ object

cannot mix interval lengths when row-binding
```

```
removed 1037 (32.17%) of rows due to 'NA' values

multiple units detected: Uhr (66.16%), units/hr (33.84%)
```

```
R> ins_treat <- ins_treat[get(index_var(ins_treat)) <= 24L]
R>
R> ins_treat <- ins_treat[,
+   list(ins_sum = .bincode(sum(ins), breaks = c(-Inf, ins_breaks, Inf)) - 1),
+   by = c(id_vars(ins_treat))
+ ]
```

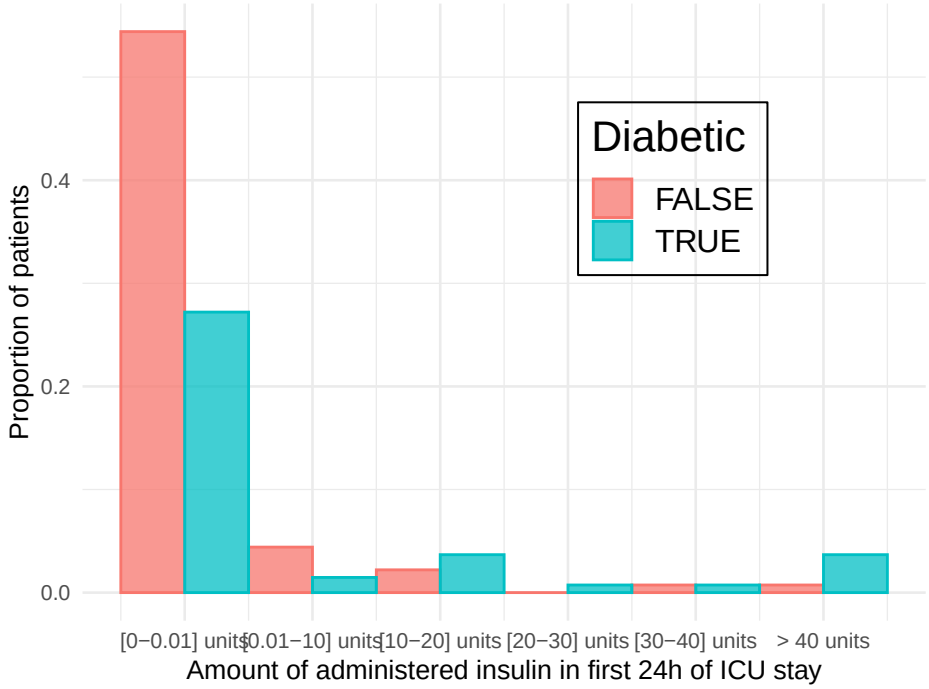
```
x does not contain column 'charttime'
```

```
R> cohort <- merge(cohort, ins_treat, by = id_vars(cohort), all.x = TRUE)
R> cohort[, Diabetic := get(id_vars(cohort)) %in% diabetes(source)]
```

```
No overlap between configured id var options and available columns for table
'diagnoses_icd' in mimic_demo
```

```
R> cohort[is.na(ins_sum), "ins_sum"] <- 0
```

After this, we can visualize the difference between the two groups with a histogram:



The plot might suggest that diabetic patients do receive more insulin that non-diabetic patients, in the first day of ICU stay.

References

Adibuzzaman M, Musselman K, Johnson A, Brown P, Pitluk Z, Grama A (2016). "Closing the data loop: An integrated open access analysis platform for the mimic database." In *2016 Computing in Cardiology Conference (CinC)*, pp. 137–140. IEEE.

Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, Shimabukuro D, Chettipally U, Feldman MD, Barton C, *et al.* (2016). "Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach." *JMIR medical informatics*, **4**(3), e28.

Evans R (2016). "Electronic health records: then, now, and in the future." *Yearbook of medical informatics*, **25**(S 01), S48–S61.

Faltys M (2018). "HIRID data." *?*, **5**, 180178.

Fleuren LM, Klausch TL, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, Swart EL, Girbes AR, Thoral P, Ercole A, *et al.* (2019). "Machine Learning for the Prediction of Sepsis, a Systematic Review and Meta-Analysis of Diagnostic Test Accuracy."

Futoma J, Hariharan S, Sendak M, Brajer N, Clement M, Bedoya A, O'Brien C, Heller K (2017). "An improved multi-output gaussian process rnn with real-time validation for early sepsis detection." *arXiv preprint arXiv:1708.05894*.

Haas SA, Lange T, Saugel B, Petzoldt M, Fuhrmann V, Metschke M, Kluge S (2016). "Severe hyperlactatemia, lactate clearance and mortality in unselected critically ill patients." *Intensive care medicine*, **42**(2), 202–210.

Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y (2017). "Artificial intelligence in healthcare: past, present and future." *Stroke and vascular neurology*, **2**(4), 230–243.

Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG (2016). "MIMIC-III, a freely accessible critical care database." *Scientific data*, **3**, 160035.

Kam HJ, Kim HY (2017). "Learning representations for the early detection of sepsis with deep neural networks." *Computers in biology and medicine*, **89**, 248–255.

Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG (2018). "An interpretable machine learning model for accurate prediction of sepsis in the ICU." *Critical care medicine*, **46**(4), 547–553.

Nichol A, Bailey M, Egi M, Pettila V, French C, Stachowski E, Reade MC, Cooper DJ, Bellomo R (2011). "Dynamic lactate indices as predictors of outcome in critically ill patients." *Critical Care*, **15**(5), R242.

Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O (2018). "The eICU Collaborative Research Database, a freely available multi-center database for critical care research." *Scientific data*, **5**, 180178.

Therneau TM, Lumley T (2015). "Package 'survival'." *R Top Doc*, **128**, 112.

Van Beest PA, Brander L, Jansen SP, Rommes JH, Kuiper MA, Spronk PE (2013). "Cumulative lactate and hospital mortality in ICU patients." *Annals of intensive care*, **3**(1), 6.

Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart C, Suter P, Thijs LG (1996). "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure."

Wang S, McDermott MB, Chauhan G, Ghassemi M, Hughes MC, Naumann T (2020). "Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii." In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 222–235.

Affiliation:

Nicolas Bennett
ETH Zürich
Seminar for Statistics Rämistrasse 101 CH-8092 Zurich
E-mail: nicolas.bennett@stat.math.ethz.ch

Drago PLecko
ETH Zürich
Seminar for Statistics Rämistrasse 101 CH-8092 Zurich
E-mail: drago.plecko@stat.math.ethz.ch

Ida-Fong Ukor
East Kent Hospitals
NHS University Foundation Trust William Harvey Hospital Kennington Road, Willesborough
Ashford TN24 0LZ
E-mail: idafong.ukor@nhs.net

Journal of Statistical Software
published by the Foundation for Open Access Statistics
MMMMMM YYYY, Volume VV, Issue II
doi:10.18637/jss.v000.i00

http://www.jstatsoft.org/
http://www.foastat.org/
Submitted: yyyy-mm-dd
Accepted: yyyy-mm-dd