The Journal of Open Source Software

Critical care data processing tools

Sinan Shi[1], **David Pérez-Suárez**[1], **Steve Harris**[2], **Niall MacCallum**[2], **David Brealey**[2], **Mervyn Singer**[2], **and James Hetherington**[1]

1 Research Software Development Group, Research IT Services, University College London **2** University College London Hospitals

Summary

cleanEHR (Shi et al. 2017) is a data cleaning and wrangling platform which works with the Critical Care Health Informatics Collaborative (CCHIC) database. CCHIC collects and gathers high resolution longitudinal patient record from critical care units at Cambridge, Guys/Kings/St Thomas', Imperial, Oxford, UCL Hospitals.

The increased adoption of high resolution longitudinal EHRs has created novel opportunities for researchers, clinicians and data scientists to access large, enriched patient databases (Harrison, Brady, and Rowan 2004) (Johnson et al. 2016). The purpose of cleanEHR is to enable researchers to answer clinical questions that are important to patients. cleanEHR is a solution to address various data reliability and accessibility problems as well. It provides a platform that enables data manipulation, transformation, reduction, cleaning and validation with a friendly user interface which empowers non-programmers to conduct basic data analysis by simply writing a human-readable configuration file.

High resolution longitudinal EHR: CCHIC

CCHIC database has in total collected 22,628 admissions (18,074 unique patients) from 2014 to 2017. It contains 119 million data points (mean 6626 data points per patient). The recruited patients have an age range from 18 to 116 years old. Physiological, laboratory, drugs and nursing information are the longitudinal data recorded during a patient's stay of the ICU. The full list of longitudinal data collected by CCHIC is listed below.

Data cleaning and wrangling

Data of this kind, though provides vast information, often faces two main issues, a) low data quality, b) low accessibility due to the complexity. We proposed a workflow, which has been incorporated in the cleanEHR package, to address the these issues. The highlight of this workflow includes the following,

- A table structure (ccTable) for data manipulation.
- Configuration file for researchers without technical knowledge to select and clean the data. The data cleaning includes various filters and data interpolation (impute) function.

For detail description of the functions and examples, please see the manual and the vignettes of cleanEHR (Shi et al. 2017)

Figure 1: List of CCHIC longitudinal data fields

References

Harrison, David A, Anthony R Brady, and Kathy Rowan. 2004. "Case mix, outcome and length of stay for admissions to adult, general critical care units in England, Wales and Northern Ireland: the Intensive Care National Audit and Research Centre Case Mix Programme Database." *Critical Care* 9 (Suppl 3). Springer Nature:S1. https://doi.org/10.1186/cc3745.

Johnson, Alistair E.W., Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. "MIMIC-III, a Freely Accessible Critical Care Database." *Scientific Data* 3 (May). Springer Nature:160035. https://doi.org/10.1038/sdata.2016.35.

Shi, Sinan, David Pérez-Suárez, Spiros Denaxas, and Steve Harris. 2017. "CC-Hic/Ccdata V1.0." https://doi.org/10.5281/zenodo.1117244.
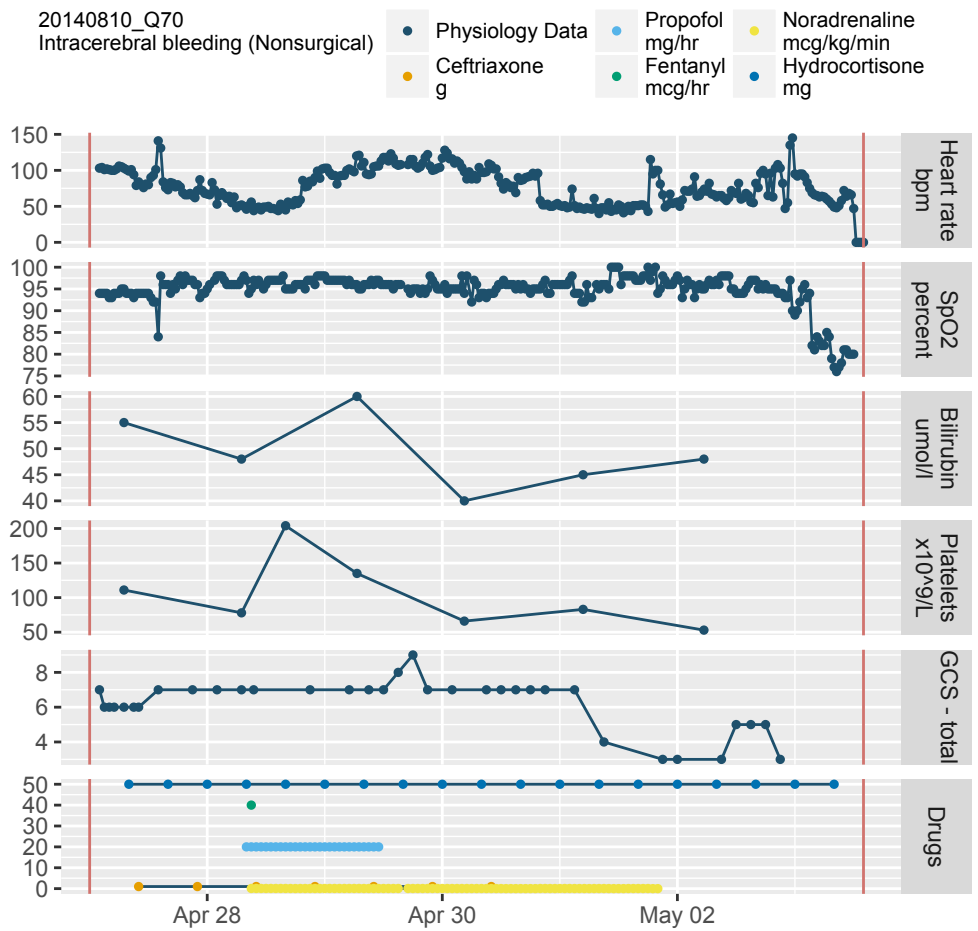
Figure 2: Selected data fields of an admission

Figure 3: An example of filtering abnormal heart rate values by range