ICUnity: A software tool to harmonise the MIMIC-III and AmsterdamUMCdb databases

Emma Rocheteau[1,2], Jacob Deasy[2], Luca Filipe Roggeveen[3], Ari Ercole[4]
[1] School of Clinical Medicine, University of Cambridge
[2] Department of Computer Science and Technology, University of Cambridge
[3] Amsterdam University Medical Centre
[4] Division of Anaesthesia, Department of Medicine, University of Cambridge

Background. The last decade has seen a proliferation of EHR databases that are set to underpin the future of healthcare. As intensive care units (ICUs) are the most data-dense in the hospital, they are particularly suited to big-data applications such as machine learning. This has led to the most well-known EHR database, MIMIC-III, which comprises data for over forty thousand patients who stayed in the critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Subsequently, this example has been emulated with the AmsterdamUMCdb database which contains 23,106 admissions of 20,109 unique patients admitted to the ICU of an Amsterdam hospital between 2003 and 2016.

Researchers working with Electronic Health Record (EHR) databases such as MIMIC-III, eICU, and the new AmsterdamUMCdb will be familiar with the challenges posed by lack of standardisation between datasets. Intra-database discrepancies include labelling of variables ("HR", "heart rate", "pulse"), units ("bpm'', "Hz''), different recording equipment (ventilator settings), and hospital standards and protocols. Moreover, extending datasets across countries leads to inter-database discrepancies, including languages ("hartslag" vs. "heart rate") and epidemiological variation in the patient population. It is likely that currently working models do not generalise to the international stage, and may not even be applicable elsewhere. Therefore, it is often necessary to test on multiple datasets to validate the robustness of a new model. Typically this would involve a huge amount of work to hand-pick features that exist in each database, and share comparable distributions.

Methods. ICUnity is a general framework for the integration of data across global EHR datasets. We focus on the MIMIC-III and AmsterdamUMCdb databases because they are extensively used by the community and they span both continents and languages. Thus far, we have developed a tool that will automatically generate likely matches between variables based on their names and data distributions. Variable strings that are in Dutch are first translated into English using the Google Translate API, before being compared to the MIMIC-III strings. String similarity is ranked using the Levenshtein distance. The data distribution of likely string matches is then compared to verify the likelihood of the match. So far, we have implemented comparison statistics based on t-tests and degree of overlap in the interquartile range. The latter is better for variables that are not normally distributed. Finally, the matches are presented to a clinician or researcher via an interactive tool for verification. The clinician can accept or reject matches, which are stored as variable links. These can then be used to generate a harmonised dataset file.

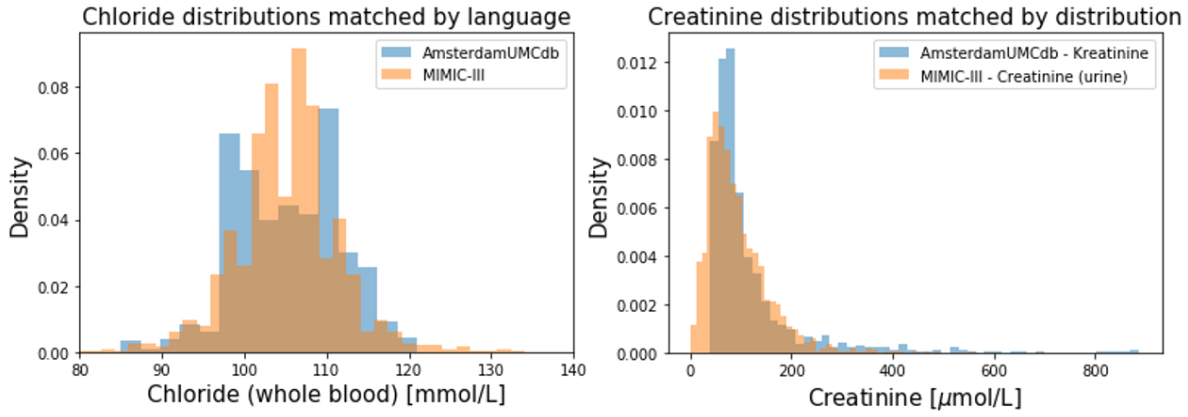

Figure 1: Example matched variable distributions across datasets. On the left: "Chloride" has a perfect linguistic match, with similar distributions. On the right: "Creatinine (urine)" and "Kreatinine" demonstrate an imperfect match which still goes on to be identified by our pipeline.

Conclusion. Development on this project begun at the Milan Critical Care Datathon 2020 and is ongoing. We welcome advice and contributions from the community. In further work, we hope to take an active learning approach. For example, if a clinician informs the system that "Calcium ion" and "Calcium (urine)" are not the same variable, then the probability of a match between "Kreatinine" and "Creatinine (urine)" will be lowered, saving time for researchers while maintaining automation. We would also like to incorporate the eICU dataset, improve

documentation of the tool, and add outlier detection features. In future, calls to translation APIs could be replaced with specialist medical-language translators, fine-tuned from existing successful translation models (e.g. GPT-2).

Link to software. The software used for this project can be found here: https://github.com/EmmaRocheteau/ICUnity