ICUnity: A software tool to harmonise medical databases

Emma Rocheteau[1], Jacob Deasy[1], Luca Filipe Roggeveen[2], Ari Ercole[1]
[1] University of Cambridge, UK; {ecr38, jd645, ae105}@cam.ac.uk
[2] Amsterdam University Medical Centre; l.roggeveen@amsterdamumc.nl

Summary

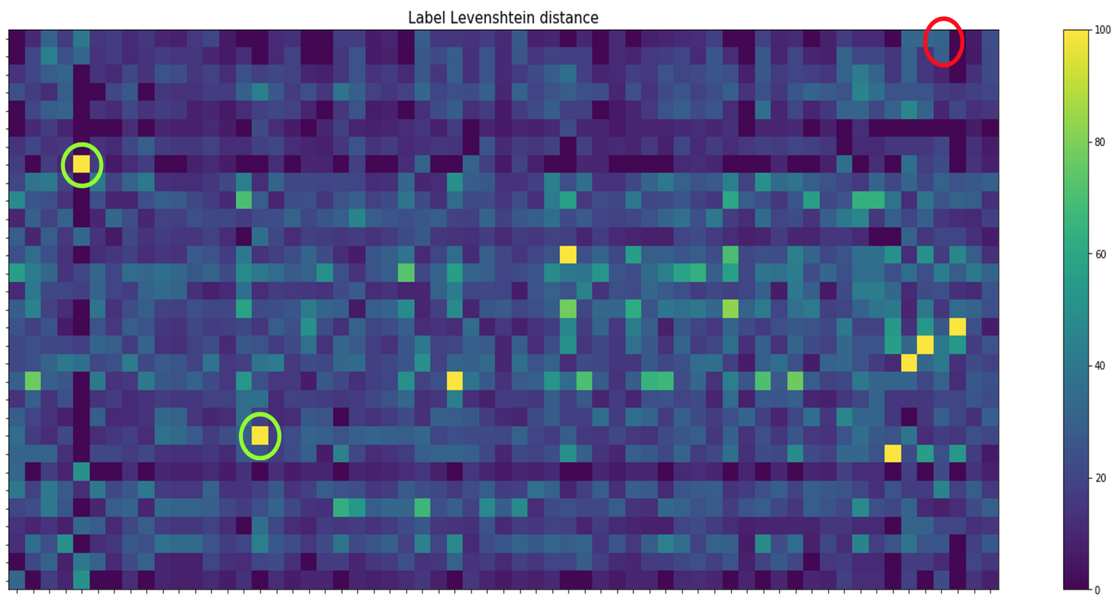
➢ There are many challenges when working with Electronic Health Record data.

➢ Typically, we would like to test our machine learning models on more than one dataset.

➢ We have developed a tool to match databases based on both string similarity and data distribution. We focus on the MIMIC-III and AmsterdamUMCdb databases.

String Similarity

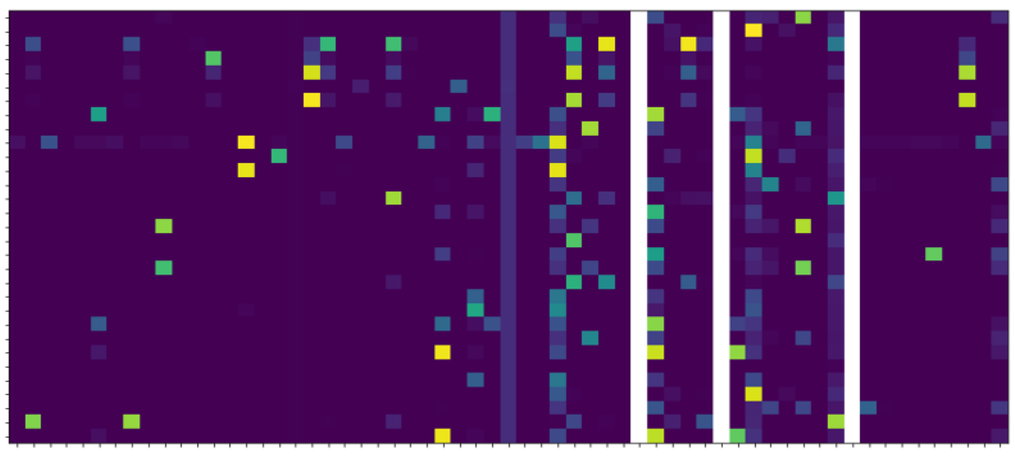
➢ Variable strings in Dutch are translated into English and then compared to the MIMIC-III strings according to the Levenshtein distance.



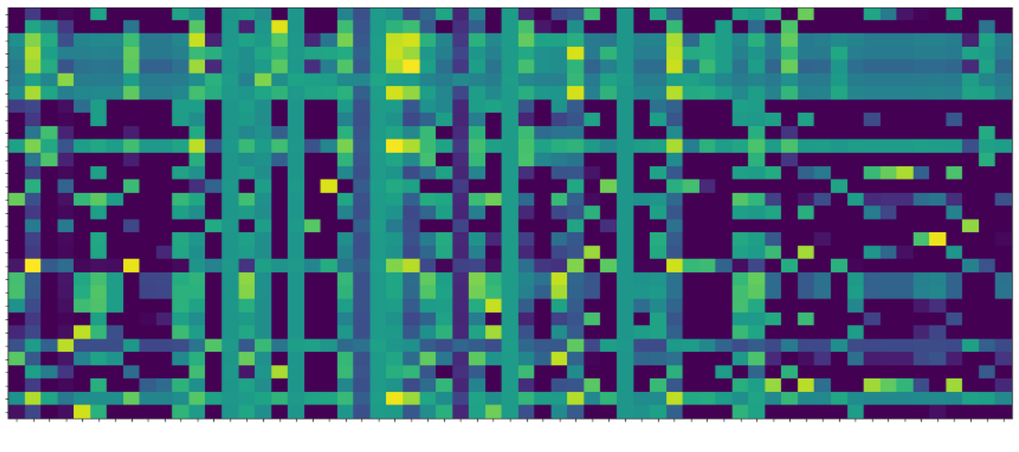
➢ This works well for some variables e.g. "Glucose" & "Glucose" and "pH" & "PH" (green circles), but not for others e.g. "PT" and "prothrombin time" (red circle).

Data Distribution

➢ We also matched based on data distributions.
➢ t-tests:



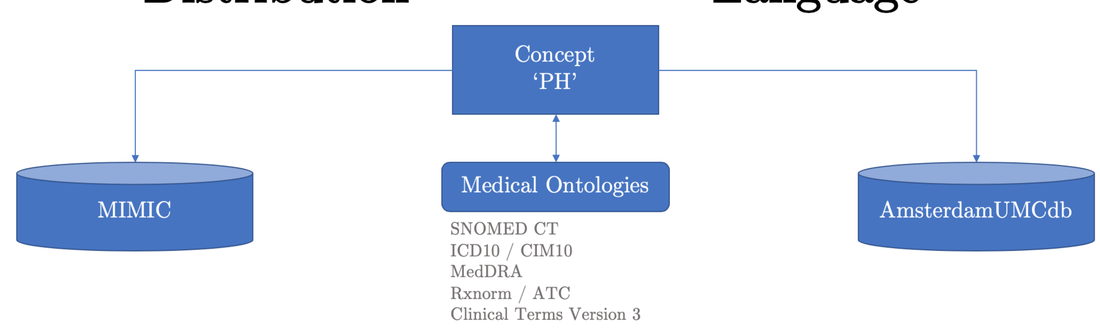
➢ Interquartile range overlap:



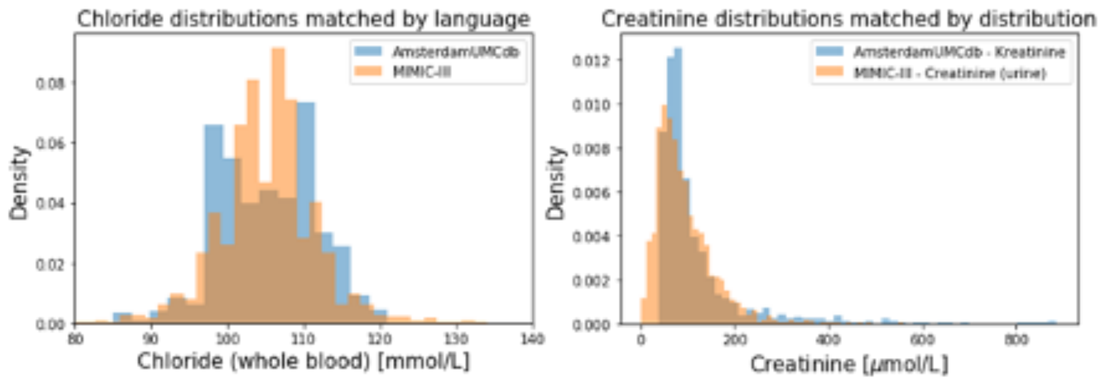
Specific challenges

Problems	Examples
Different formats	MIMIC, eICU, AmsterdamUMCdb
Multiple & different label IDs	“HR”, “heartrate”, “pulse”
Different languages	“Hartslag” vs. “heartrate”
Different units	“Mmol/l” vs. “g/l”
Different sources	Blood pressure: femoral, radial
Numeric vs. string	5 vs. “>2”
Granularity	“20-30y” vs. “20”, “21”...
Different distributions	Different populations

Data Harmonisation



Examples



Software Tool

➢ We have a basic validation tool for clinicians to confirm or deny the suggested matches.

```
Is this a suitable match:  Kreatinine, Creatinine (serum)
Confirm? [Y/N] or X to exit: yes

Is this a suitable match:  Chloor, Chloride (urine)
Confirm? [Y/N] or X to exit: n

Is this a suitable match:  Kreatinine, Creatinine (pleural)
Confirm? [Y/N] or X to exit: n

Is this a suitable match:  Kreatinine, Creatinine (ascites)
Confirm? [Y/N] or X to exit: no

Is this a suitable match:  Albumine, Albumin (urine)
Confirm? [Y/N] or X to exit: exit
```

```
[28] print('Saved Matches:')
     print(matches)
     print('Non Matches:')
     print(non_matches)
```

```
Saved Matches:
[['PH', 'pH'], ['Calcium', 'Calcium'], ['Natrium', 'Sodium'], ['Kalium', 'Potassium'], ['Fosfaat', 'Phosphate'], ['Kreat
Non Matches:
[['Calcium ion', 'Calcium (urine)'], ['Calcium ion', 'Calcium'], ['Kreatinine', 'Creatinine (urine)'], ['Chloor', 'Chlor
```

Conclusion

➢ We have started development on a tool to harmonise EHR databases.

➢ In future, we would like to incorporate more sophisticated matching e.g. using active learning to reduce mistakes.

➢ GitHub link: https://github.com/EmmaRocheteau/ICUnity.