COMP90089: Machine Learning Applications for Health

Semester 2, 2023

Brian Chapman, PhD; Daniel Capurro, MD, PhD

Assignment 1: Digital Phenotypes

One key step in most machine learning projects involves identifying the group of patients that will be included in the study (we call this a patient cohort) and the labels that you might be interested in. Most of the time you will NOT find a dataset that is fully curated and clean, individual records will not have the labels that you want. How do we know which patients to include? How do we label them?

Clinical databases store information that–ideally–represents what happens in real life. However, clinical databases that were not collected for research purposes have significant data quality issues. The fact that a patient has a diagnostic code (e.g. ICD codes) does not mean that the patient actually has the condition so it does not work (by itself) as a way to define a cohort or as a label. We usually need to construct digital phenotypes. Digital phenotypes are combination of data elements that accurately define a patient characteristic. You can find a good example here
For this first assignment, you will propose a digital phenotype for a relevant clinical condition in critical care and describe how you would create a cohort of such patients using MIMIC.

1. Read the introductory Up To Date chapter on hypotension and shock here. You should have access through the University of Melbourne Libraries.

2. List the criteria used by clinicians to define **Hypotension**.

3. List the International Classification of Diseases (ICD) codes used to describe hypotension. You can use the World Health Organisation's code browser.

4. SNOMED-CT is a clinical ontology. Use CSIRO's Shrimp browser to navigate SNOMED-CT and find the concept that would match the ICD-10 code for hypotension. Is there a perfect match? Please describe the potential pros and cons of mapping between one terminology (ICD codes) and another (SNOMED-CT) when doing a data analysis project that requires linking data from more than one source.

5. Study the MIMIC data model. Describe your algorithm to find patients that meet this definition of profound hypotension:

   - Hypotension that lasted $>$ 60 minutes and was corrected after the use of intravenous vasoactive drugs.

   Describe your algorithm **in plain English** and implement it **in SQL and Python**. Submit your implementation as a *Jupyter notebook*.

   Document the rationale behind your implementation in the notebook with a sufficient level of detail (including any assumptions). Your **final cohort table** should

include the patient identifiers and remaining columns used for filtering. For each patient that was picked up by your algorithm, include a column that states whether ANY ICD code for hypotension was added to their list of diagnoses at the time they were discharged (hint: diagnoses_icd table).