Datasets for Workshop

The GRAPH Courses team

May 2023

Workshop Datasets

In this week's workshop, you will analyse a dataset of your choice. These datasets can be found in the "data" subfolder.

The scripts with instructions for analyzing the datasets are in the "scripts" subfolder.

1. Febrile Diseases Burkina Faso

This dataset focuses on febrile diseases in Burkina Faso, including clinical data from 914 hospital patients suffering from fever-causing illnesses. The researchers aimed to test the accuracy of an algorithm that utilizes blood metric measurements to predict the cause of infection, comparing gold standard culture techniques with a haemocytometry-based algorithm (IMS), procalcitonin (PCT), and C-reactive protein (CRP). The dataset provides information on patient demographics, symptoms, and blood test results, as well as the final classification of the infection cause under the variable ims_final_full_classification.

Find out more about the dataset **here**.

2. India Tuberculosis Treatments

The India TB Pathways dataset is based on a study conducted in Vellore district, Tamil Nadu, India. It examines the pre-diagnostic costs and health-seeking behavior of 880 new adult pulmonary tuberculosis patients registered at Primary Health Centres (PHCs). The dataset reveals that a majority of patients initially visited private health facilities, resulting in substantial out-of-pocket (OOP) medical costs and delays prior to diagnosis and treatment initiation.

Find out more about the dataset **here**.

3. Motorcycle Accidents in Colombia

This dataset contains police records of deaths resulting from motorcycle-related road injuries in the city of Medellin between 2012 and 2015. It focuses exclusively on fatal incidents involving motorcyclists, both drivers and passengers. The relevant dataset at the linked site is called "Dataset 1.xlsx". In the second sheet of that xlsx file, you will find a data dictionary. The accompanying publication can be accessed **here**.

Access related data **here**.

4. COVID-19 Patients in the US

This dataset encompasses anonymized data from 1540 COVID-19 hospitalized patients at two New York hospitals: State University of New York (SUNY) Downstate Health Sciences University and Maimonides Medical Center. The data, found in the "demographics_both_hospitals.csv" file, provides a wealth of information on patient demographics, known comorbidities, and ultimate hospitalization outcomes (discharge or death).

Find out more about the data **here**.

5. Sex Attitudes Survey in Britain

This dataset contains anonymized data from the National Survey of Sexual Attitudes and Lifestyles, third wave (Natsal-3), one of the largest studies of sexual behaviour in Britain. Conducted between 2010 and 2012, Natsal-3 interviewed over 15,000 British adults aged 16 to 74. The data consists of responses from 3,799 adults aged 16-74. The variables cover respondents' demographic characteristics, attitudes towards sex and relationships, and reproductive history.

Find out more about the data **here**.

The variable dictionary can be found **here**