Filters applied

This document provides an overview of the preprocession steps took to assemble the dat datasets compiled from the Chikungunya disease database. For more detailed information about the datasets, including data collection methodologies, data sources, and data preprocessing steps, please visit the project repository on GitHub: `https://github.com/gabivaleriano/HealthDataBR`.

General preprocessing

Filter	**Size**
Initial dataset	269960
Filter 1: remove duplicates	269823
Filter 2: remove if is na for state or health unit	269768
Filter 3: filter only confirmed positive and negative cases of the disease	237764
Filter 4: remove patients without year of birth	235845
Filter 5: remove patients older than 110 and younger than 12 years old	213701
Filter 6: remove patients without sex information	213483
Filter 7: remove patients without race information	183797
Filter 8: remove patients without schooling information [1]	183693
Filter 9: keep only patients in the acute stage of the disease	138724
Delete columns without relevant information, filled mostly with NA's or redundant	

Table 1: General preprocessing steps.

Preprocessing for the *hospitalization* dataset

Filter	**Size**
Initial size	138724
Filter 1.1: keep only patients with the disease confirmed	107325
Filter 1.2: keep only patients with information about hospitalization	71905
Filter 1.3: remove patients hospitalized more than 15 days after the first symptoms	71782
Filter 1.4: remove non-hospitalized patients that died	69686
Delete columns without relevant information	
Number of hospitalized patients	1681

Table 2: Additional preprocessing steps for the *hospitalization* dataset.

Preprocessing the for *diagnose* dataset

Filter	**Size**
Initial size	138724
Delete columns without relevant information	
Number of positive cases	107325

Table 3: Additional preprocessing steps for the *diagnose* dataset.

Preprocessing the for the *death* dataset

Filter	**Size**
Initial size	138724
Filter 3.1: only confirmed cases of the disease	107325
Filter 3.2: only patients cured or dead with the disease	103176
Filter 3.3: keep only hospitalized patients	1619
Filter 3.4: remove patients that died with more than 30 days after hospital admission	1611
Delete columns without relevant information	
Number of death cases	44

Table 4: Additional preprocessing steps for the *death* dataset