**Health system performance during Covid**

**Data cleaning protocol**

Last updated: November 11, 2020

Before uploading the data to the online dashboard and performing analyses, we will clean the data, assess completeness and restrict the sample to health facilities with “acceptable” reporting over the period of interest. These methods have been implemented in STATA. The code used is available in the project GitHub at: <https://github.com/catherine-arsenault/HS-performance-during-covid-do-files> Cleaning do files are called clean\_ETH, clean\_HTI, clean\_KZN, clean\_NEP etc.

1. **Missing values:** An important limitation of DHIS2 data relates to missing values. When there are no patients for a specific service or no deaths for a specific indicator, a missing value is often found instead of a 0. For volume of services (e.g. number of sick child visits), we cannot tell if a missing value means 0 patients or if that the facility did not report that month. Missingness is very frequent for mortality data (e.g. stillbirths or newborn deaths). We decided to impute 0s for mortality if the service that it relates to was provided that month. The following imputations were made:
   1. Stillbirths: missingness replaced with 0 if there were deliveries conducted in the facility that month
   2. Newborn deaths: missingness replaced with 0 if there were deliveries conducted in the facility that month
   3. Maternal deaths: missingness replaced with 0 if there were deliveries conducted in the facility that month
   4. Inpatient deaths: missingness replaced with 0 if there were inpatient admissions in the facility that month
   5. Emergency department deaths: missingness replaced with 0 if there were emergency room visits in the facility that month
   6. ICU deaths: missingness replaced with 0 if there were ICU admissions in the facility that month
2. **Outliers:** Because dhis2 data is self-reported by health facilities some errors can be found in the data (e.g. typos, impossible values entered etc.) For every variable, we will assess the presence of extreme positive outliers. WHO1 recommends the use of multiples of the standard deviation of the mean to identify outliers. For each health facility, values that are greater than 3.5 SD from the mean will be considered as extreme positive outliers. We only assess for positive outliers (i.e. extremely large values) and not for negative outliers (extremely small values) because we are expecting volumes of services to decrease substantially after Covid. We will also only assess outliers if the mean over the period is greater than one. This technique avoids flagging as outlier a value of 1 if the facility reports: 0 0 0 0 0 1 0 0 0 0 0 0 which is common for mortality data.
3. **Complete case analysis:** In most countries, completeness of reporting is an issue. Certain facilities do not report every month and certain facilities have very long reporting delays. In order to measure changes in volumes of services over time, we need to include only health facilities with consistent reporting. However, in some countries, if we were to include only health facilities that report every indicator every month, we would exclude a lot of health facilities and end up with a very small (and non-representative) sample of health facilities. We therefore opted to create two datasets:
   1. One dataset keeps all facilities that reported nearly all periods (e.g. 14/18 months or 16/20 months). This will be used for the dashboard.
   2. One dataset keeps facilities that reported the latest months of interest e.g. April, May, June 2020 and the matching months in 2019 (April, May June 2019). This dataset will be used for simple calculations contrasting Q2 2020 to Q2 2019.