Preparing the data collection

Starting on the 19th of April 2020, trained enumerators started collecting data on the main health facilities of the country. These enumerators had some level of experience as they had all previously participated in Integrity Watch's Community Based Monitoring Programs. Before going to the field, they all participated in a two-day intensive training covering the use of the technology for data collection, the questionnaire in detail and the methods of neutrality when asking questions, as well as physical and health security protocols.

Given the current COVID-19 situation, this last part was given special attention, focusing on personal hygiene, specific material to be used, and distance. All enumerators went to the field with hand cleaning solutions (>70% alcohol), with gloves that they changed after each visit, masks, phone pockets regularly cleaned with alcohol wipes, and a thermometer to check temperature twice a day. Enumerators were also briefed on how to properly wash hands and to do so as much as possible, especially in between activities, after touching phones, coughing, or sneezing, after interactions, etc. If any of them felt a headache, started coughing or had a temperature above 37.5°C they immediately stopped working and quarantined themself.

The data collection process has been done in a collaboration between Integrity Watch Afghanistan and VoxMapp, a social-tech company. Integrity Watch provided the enumerators, facilities for training, and field coordination, while VoxMapp provided the mobile app for data collection, monitoring of activities, data cleaning, analysis, and visualization. The questionnaire design and training have been created by the two organizations together.

The data

The first layer of data collected is what we call a baseline. The idea is to scan the main countries' health facilities to form a database with 1) general information on each health facility and 2) a first layer of information related to COVID-19. This general information includes basic information as well as measures of capacity and readiness to face strenuous situations, such as a pandemic. We thus collect info on: GPS location, name of the health facility, type of health facility, name and contact of hospital director, date of construction, funding of the construction and management, measures of overall conditions (infrastructure condition, access to water, electricity and toilets), measures of capacity (number of doctors, nurses, beds, daily capacity on normal days, average number of weekly patients, etc). This general information will serve as a first layer of data on the health facility, that will also be useful to perform calculations such as measuring hospital's saturation. The data collected on COVID-19 concerns two main aspects: readiness to face and spread of the disease. This data is collected first during the baseline, and then is updated daily by the health facilities using VoxCovid, a mobile app and online software designed by VoxMapp specifically to tackle COVID-19.

Each infrastructure is given a unique id number and password that allow them to update information related to their health facility only (through the mobile app or online software). This information includes: screening procedures; awareness campaigns; test kits; availability of protective equipment for its personnel (alcoholic solutions, soap and water, P2 masks, vinyl gloves, visors, aprons, shoe covers, protection hats); intensive care machines (respirator, and x-ray); oxygen, medicine and anaesthetic reserves; number of patients showing COVID-19 symptoms; number of patients testing positive for COVID-19; number of patients in intensive care; number of deaths from the virus; number of recovered from the virus; number of deaths that are not from COVID-19.

Ensuring data quality

Ensuring the highest quality of our data is at the core of our work. There are many layers of controls and quality checks.

The first layer is built-in our survey tools: the questionnaire, the mobile app, and the online software. The questionnaire is designed and tested to ensure that questions are formulated in a comprehensive, delimited way, with clear boundaries and easy to estimate for the interviewee. We also triangulate some information to cross-check responses. For instance, we not only ask the number of deaths from COVID-19 yesterday, but we also ask for that same number since last Friday. Questions related to COVID-19 cases are all asked remitting to the day before, to ensure that the numbers will not further change during the course of the day. Specific instructions are given in the questionnaire to the enumerator, to avoid any confusions. The digital tools ensure data quality and completeness in the following ways: they do not allow for answers to be left in blank, numerical answers are entered by scrolling through numbers and not by typing, and they save information provided in offline settings to be later uploaded to the base. The digital tools also allow us to directly communicate with the enumerators if needed, and to set-up polygons of delimited areas to avoid data overlapping.

Another layer of quality comes with ensuring proper training of the enumerators. This includes training on understanding of the questionnaire, social skills on question asking, and technical skills on the use of the app. Each enumerator gets some practice not only in the training, but also in the field.

One of the most important steps to ensure the quality of our data is the daily monitoring that is performed by VoxMapp analysts. This includes checking for wrong contact numbers, cross logical conflicts, interview duration and more generally a set of high frequency checks. Everyday a sub-database is extracted containing potential problematic entries that need some sort of revision. In the most pressing cases the health facilities are contacted via phone to verify the information.

Finally, to cross-check the daily information that is given by the health facilities directly, Community-Based Monitors from Integrity Watch are assigned to monitor a specific area containing multiple health facilities. Once a week, they will visit health facilities to verify to information previously provided.

Analysis and publication

After daily cleaning and analysis, the data is being published in the form of daily updated dashboards.

The first batch of dashboards maps the urgency of needs at three different levels: country level, provincial level, and health facility level. Questions about equipment reserves are asked in terms of “days left” of each material/equipment at the health facility level. The categories are “Not available”, “Available for 1 day”, “Available for 3 days”, “Available for 7 days”, “Available for at least 15 days”. These variables are transformed into numerical variables, for us to perform calculations. We generate numerical variables for each category of personnel equipment and medicines, as well as an average measure that defines the level of urgency at the provincial and health facility level. This global measure allows us to spot the provinces, and even health facilities that need more pressing attention in terms of materials distribution. The dashboard also shows a colour from green shades to red shades that defines the level of urgency, where availability for 15 days is dark green, and for 0 days is dark red. The centre is skewed to the upper values, so that the lighter red starts appearing when it is appropriate to start planning action (i.e. 10 days).

We also generate calculations to assess the need for more test kits and for more respiratory machines. For the test kits we calculate the percentage of symptomatic patients with access to test kits, that is,

$$Access\ to\ test\ kits_{ht} = \left(\frac{Number\ of\ available\ test\ kits_{ht}}{Number\ of\ symptomatic\ patients_{ht-1}}\right) * 100\ ,$$

where *h* is the health facility and *t* is today.

For the respiratory machines we calculate the percentage of intensive care patients with access to a respiratory machine, that is,

$$Access\ to\ respiratory\ machines_{ht} = \left(\frac{Number\ of\ available\ respiratory\ machines_{ht}}{Number\ of\ intensive\ care\ patients_{ht-1}}\right) * 100\ ,$$

where *h* is the health facility and *t* is today.

In the dashboard these two values will turn to red whenever they are below 100 percent and green if above, because any percentage below 100 indicates a shortage of test kits or respiratory machines.

The data on the needs is presented at the country level, as an average of all health facilities. This is interesting to get an overview of the country's situation. We also present the data at the provincial level, as an average of health facilities inside each province. This is insightful in that it provides knowledge about the areas that need more urgent attention. Finally, we provide this data at the disaggregated health facility level. This level of disaggregation is especially important for decision making and for acting. It allows the government and humanitarian organizations to make informed decisions on which facilities need materials and equipment more urgently.

We are also publishing an updatable dashboard mapping the implementation of screening procedures and awareness campaigns at the health facility level. In the baseline questionnaire and daily updates generated by the health facility we ask if the health facility has a screening procedure in place to separate COVID-19 patients from other patients, as well as if there have been awareness campaigns in that hospital (including posters on the walls, etc). Both of these questions are very important: it is important to identify which health facilities are not screening patients as this could be a huge factor of spread of the pandemic; and it is also useful to know what health facilities might need trainings and materials to spread awareness.

Finally, the last batch of dashboards maps the risk areas in terms of virus spreading. We use the latest number of symptomatic, tested positive, deaths and recoveries to generate the number of active cases at the provincial level today. We are not publicly disclosing information on the specific number of COVID-19 cases and deaths as we understand this information is sensitive; we simply create and map a provincial ranking from the most affected province to the least affected province, in terms of current active cases. The formula used to calculate active cases is as follows:

$$Active\ cases_{pt} = \sum_{i=1}^{t-1} Symptomatic_{pt} + \sum_{i=1}^{t-1} Tested\ Positive_{pt} - \sum_{i=1}^{t-1} Deaths_{pt} - \sum_{i=1}^{t-1} Recoveries_{pt}\ ,$$

where *p* is province, *i* is the first day of data collection, and *t* is days since data collection began (today).

We also generate a risk indicator at the district level containing the three following parameters: average days left with material and equipment in that district, saturation of the hospitals in the district, and percentage increase in district cases.

The average days left with material and equipment is a simple average of the global needs, as discussed in the first batch of dashboards. The saturation of hospitals measure is calculated as follows:

$$Saturation_{dt} = \left(\frac{Intensive\ care\ patients_{dt-1}}{Number\ of\ beds_{d}}\right),$$

where *d* is district, and *t* is today.

And, the percentage increase in district cases:

$$Percentage\ increase_{dt} = \frac{(Symptomatic_{dt-1} + Tested\ Positive_{dt-1}) - (Symptomatic_{dt-2} + Tested\ Positive_{dt-2})}{(Symptomatic_{dt-2} + Tested\ Positive_{dt-2})},$$

where, *d* is district and *t* is today.

Each part of the indicator is calibrated as follows:

1) If the average days left with materials and equipment is between 0 and 4, it will be considered high risk; if between 5 and 10 days, medium risk; and if between 10 and 15 days, low risk.

2) If the saturation is between 75 and 100 percent (or more), it will be considered high risk; if between 25 and 75 percent, medium risk; and below 25 percent, low risk.

3) If the percentage increase in cases is between 75 and 100 percent (or more), it will be considered high risk; if between 15 and 75 percent, medium risk; and below 15 percent, low risk.

The results will then generate a comprehensive measure of risk at the district level.

Finally, we also represent graphically the *country growth factor curve of daily new cases* which is a very good measure of the exponentiality of the spread of the virus:

$$Growth\ factor_{t-1} = \frac{New\ cases_{t-1}}{New\ cases_{t-2}},$$

$$New\ cases_{t-1} = Symptomatic_{t-1} + Tested\ Positive_{t-1}$$

and,

$$New\ cases_{t-2} = Symptomatic_{t-2} + Tested\ Positive_{t-2},$$

where *t* is today.

If, Growth factor is > 1 for a sustained period of time, then the pandemic is in exponential growth.

If, 0 < Growth factor < 1, then there is a non-exponential increase in cases.

If, Growth factor < 0, then there is a decrease in cases.