POLICY NOTE NO. 6 / 08 May 2020

Prevailing Data Issues in the Time of COVID-19 and the Need for Open Data

UP COVID-19 Pandemic Response Team

Introduction

Our previous policy notes[1] already highlighted some epidemiological metrics vital for monitoring the state of the country in its fight against COVID-19. These metrics provide context, give insight, and serve as a guide to help stakeholders take control of the disease and monitor our own progress. As we have repeatedly pointed out since Day 1, any analysis is only as good as the data that we have.

Relevant, and accurate data about COVID-19 and the resources the country has in the fight is important. The speed and timeliness of how such is released is equally critical. We do recognize the difficulties in the collection of detailed and timely records of COVID-19 data on a nationwide scale. However, no matter how difficult, these problems need to be addressed at the soonest possible time and should be of top priority. Here, we discuss prevailing data issues we have observed, and our recommendation for open data moving forward.

Data Sharing Practices

Collecting COVID-19 data has not been easy. Since the country finally accepted the presence of community transmission around the first week of March 2020, government's data reporting protocols have changed far too often (Table 1). There was initially no standard time of the day for when DOH posted the official

[1] See, in particular, Policy Note No. 2 - Modified Community Quarantine beyond April 30: Analysis and Recommendations

daily numbers. There was even a brief period when DOH stopped releasing updates, leaving people using the data in the dark.

Table 1. Timeline of changes in DOH's COVID-19 data reporting

Data Reporting Changes	Dates 3/10 to 3/11	3/12	3/13	3/14	3/15 to 3/16	3/17 to 4/2	4/3 to 4/5	4/6 to 4/13	4/14 to 4/26
I. Data Reported									
1. Count of COVID-19 Positive Cases	■	■	■	■	■	■	■	■	■
2. Count of PUI	■	■	■	■	■	■	■		
3. Count of PUM	■	■	■	■	■	■			
4. Case Information									
a. Demographics									
- Age	■	■	■		■	■	■		■
- Sex	■	■	■		■	■	■		■
- Residence	■	■	■		■	■	■		■
- Nationality	■	■	■		■	■	■		
b. Health Information									
- History of travel/exposure	■	■	■		■	■	■		
- Date of Onset	■	■	■	■	■				
- Date of Admission			■	■	■				
- Date of Laboratory Confirmation	■	■	■		■	■	■		■
- Health Facility Admitted	■	■	■		■	■	■		
c. Death Information									
- Date of Death		■	■	■	■	■	■		■
- Cause		■	■	■	■	■	■		
- Co-morbidity		■	■	■	■	■	■		
d. Date of Recovery					■	■	■		■
II. Format of Case #									
1. PHx; x=1,2,.,,	■	■	■	■	■	■	■		
2. Cx; x is random									■
III. Reporting Platform									
1. Summary of Case Information - pdf format	■	■	■		■				
2. ncovtracker individual case information						■	■		
3. data drop - google drive									■

Fortunately, the DOH has been open to feedback, and this has been reflected in the improvement in reporting. There is now a centralized resource[2] for modelers to use as a common reference. There is also regularity and predictability on when reports are given, and this is commendable given that transparency and timeliness are key to managing any crisis situation, especially a pandemic.

However, there is still room for improving the quality of the data, and the process of collecting information for DOH's data drop in the tracker. It is

[2] See https://www.doh.gov.ph/covid19tracker

important to make sure that correct data is captured as swiftly as possible to minimize uploading of erroneous and anomalous, if not missing, data.

Data Accuracy and Integrity

The availability of accurate and relevant data is a basic requirement in managing any situation that requires urgent and targeted response. Almost three months since we had our first confirmed case in the person of a Chinese national on 30 January, we have yet to reconcile differences in numbers between DOH and LGU sources. For example, on 03 May 2020, DOH reported 7 deaths (28 recoveries) in Laguna, which was 22 deaths (65 recoveries) lesser than the provincial government's official count.

Accuracy, however, goes beyond correctness in reporting aggregate numbers. Recent data drops by DOH revealed a number of alarming patient-level inconsistencies, if not gross errors. A quick comparison of the April 24 and April 25 data drops showed that forty-five (45) cases have changed sex from male to female or vice-versa; while 75 others had the data on age modified. This is on top of the 516 cases where the residence data was reclassified to another city, if not a completely imaginary city (i.e. a barangay or district) like what happened in the City of Manila (Table 2).

Table 2. Example of errors in reported sex, age, and residence of confirmed cases as released by DOH on 24 April and 25 April 2020

Sex		
Case#	April 24	April 25
Cxxxx66	M	F
Cxxxx29	M	F
Cxxxx57	F	M
Cxxxx27	F	M
Cxxxx12	F	M
Cxxxx20	M	F
Cxxxx70	F	M

Age		
Case#	April 24	April 25
Cxxxx47	28	29
Cxxxx05	48	49
Cxxxx92	36	35
Cxxxx82	8	1
Cxxxx16	31	28
Cxxxx91	21	46
Cxxxx61	28	70

City/Municipality of Residence		
Case#	April 24	April 25
Cxxxx19	City of Manila	Sampaloc
Cxxxx73	City of Manila	Santa Ana
Cxxxx51	City of Manila	Santa Cruz
Cxxxx10	City of Manila	Pandacan
Cxxxx78	Quezon City	City of San Juan
Cxxxx16	City of Manila	Sampaloc
Cxxxx23	City of Manila	Santa Cruz

Cxxxx25	M	F	Cxxxx96	39	34	Cxxxx58	City of Manila	Ermita
Cxxxx78	M	F	Cxxxx29	60	28	Cxxxx08	City of Manila	Sampaloc
Total erroneous cases: 45			**Total erroneous cases: 75**			**Total erroneous cases: 516**		

Related to the problem of accurate residential reporting is the handling of certain variables in the DOH data drop. For example, RegionRes is a variable for the region of residence and is coded in text such as "NCR", "Region III: Central Luzon", and so on. RegionPSGC is the region code based on the Philippine Standard Geographic Code [PSGC], compiled by the Philippine Statistics Authority, based on their mandate to prescribe "uniform standards and classification systems in the generation of government statistics to ensure harmonization and comparability of statistics in the country and at the international level" (PSA 2020). The DOH data drop introduced this variable last April 26, 2020 for regional, provincial, and city/municipal classification. Each region has a unique code in the PSGC; however, as seen in the table below, inconsistencies in the coding of the PSGC for cases within regions have been problematic. It is noted that the DOH Tracker uses the RegionRes variable for statistics, not the Region PSGC (Table 3).

Table 3. Data inconsistencies from DOH data drop on 6 May 2020

Row Labels	Count of DateRepConf
NCR	**6596**
PH030000000	2
PH040000000	5
PH130000000	6581
PH140000000	1
(blank)	7
Region III: Central Luzon	**389**
PH030000000	383
PH130000000	6
Region IV-A: CALABARZON	**1245**
PH040000000	1237
PH130000000	7
(blank)	1

Row Labels	Count of DateRepConf
Region IV-B: MIMAROPA	**28**
PH040000000	1
PH170000000	26
(blank)	1
Region IX: Zamboanga Peninsula	**49**
PH090000000	46
(blank)	3
Region V: Bicol Region	**55**
PH050000000	48
(blank)	7
Region VII: Central Visayas	**1180**
PH070000000	1179
PH130000000	1

Grand Total:	**9542***

** Total does not necessarily agree with the 10,004 cases reported in May 6, 2020 as regions with full-matching PSGC codes have been disregarded.*

There are other troubling anomalies in recent data drops of DOH. For example, 18 cases no longer have data on residence in the April 25 update. On the same date, the recovery dates of two cases were either missing or changed. One patient who reportedly died on April 24 is no longer dead the following day.

The DOH data drop is also inconsistent with its use of date formats, which makes it difficult for automated systems of extracting and updating data from case information. It has made the work of data analysis difficult because of these sudden changes (Table 4).

Table 4. Inconsistency in reporting of date formats

Date Format	Date First used by DOH Data Drop
MM/DD/YYYY (e.g., 04/14/2020)	April 14, 2020
DD-MMM-YYYY (e.g., 27-Apr-2020)	April 22, 2020 Ø except DateRepRem (Date of report of removal, which is recovery or death), which is MM/DD/YYYY, Ø all dates in similar format by April 24, 2020
YYYY-MM-DD (e.g., 2020-04-14)	May 5, 2020

These lapses may seem small relative to the total size of data contained in the daily updates, but they have significant implications on the reliability of our scientific analyses on COVID-19. Patient case data is the keystone for effective and insightful metrics and analysis. The integrity of the data drops is particularly important given that no less than President Rodrigo Roa Duterte himself has said many times that the government's decision on managing COVID-19 will be based on science. We fully support President Duterte on this call for science-based decisions, hence this statement.

Transparency and Accountability

We acknowledge the importance of data privacy as provided for in our existing laws such as the Data Privacy of 2012 (RA 10173) and the Mandatory Reporting of Notifiable Diseases and Health Events of Public Health Concern Act of 2018 (RA 11332), among others. However, there are important data that can already

be anonymized and made available to serve public interest. For example, identifiers, such as employment information or specific addresses may be removed, but variables such as onset of symptoms, exposure history, co-morbidities, and whether they were medical front-liners or not are key inputs for modelers and statisticians to map the progress of our fight against COVID-19.
We are also aware that the DOH is already sharing government data with selected groups from the private sector. These organizations are bound by Non-Disclosure Agreements (NDA), as required by law. However, while it may be legally right, it does not serve public interest in this time of great need for accurate and timely information. For example, DOH restricts the analytics involved with patient statistics, even for some aggregates, which have implications on understanding IATF's recommendations for placing some provinces under ECQ.

The COVID-19 pandemic requires a science-based approach, and science cannot exist in a vacuum. Any scientific output would benefit from cross-validation from peers, and if findings do not converge, we might be standing on shaky grounds. Such scientific rigor can only happen in an environment where data, especially government data, is made available to all relevant stakeholders. Entrusting government data to select private entities is inimical to public interest.

Call for Open Data and Scientific Cooperation

We understand that some data can only be shared internally (i.e. within the government) and are not fully open to the public. In this regard, we call on other agencies, to share relevant data that can help capable institutions make scientific assessments for discussions on the evolving crisis to come up with better peer-reviewed science. Regardless of technology, it is important that the reporting system be standardized and regularized, integrated into the existing data tracker as much as possible, and made open to the public.

We also call on private institutions to contribute to the COVID-19 related data already shared in the COVID-tracker Data Drop. We believe there are private corporations who possess data that can benefit all researchers cooperating in this fight.

Making all data sources open, while also being mindful of the same data privacy protocols that DOH is following, can further empower both official and

unofficial stakeholders (i.e., commissioned and independent scientists and researchers, local governments officials, IATF/NTF decision-makers) in the battle ahead. This is important not only to inform our plans, but also to tell us how we are doing in the fight against COVID-19. This is particularly so in light of recent announcements by DOH and its private partners that we have already flattened the curve. Without access to full government data entrusted to select private sector groups, the task for an independent corroboration—the hallmark of any scientific undertaking—becomes impossible, to the detriment of public welfare and interest.

This call for open data is in line with the UNESCO call for open science and reinforced scientific cooperation. According to UNESCO, it is imperative now more than ever to strengthen/build international inter-continental and national scientific cooperation between scientists, decision/policy makers, private practitioners, industries and health professionals and civil society for a multi-dimensional approach to tackling the pandemic. This calls for open access to scientific knowledge and know-how, data sharing and evidence-based policy and decision-making.

Nowhere is the need for Open Data as clearly manifested than in the current COVID-19 crisis. In preparing for, responding to, and recovering from the impacts of health hazards or any natural hazard for that matter, data must be used to generate knowledge. If we keep our information in silos, our collective efforts and perspective of the situation narrows and so do our chances to maintain and preserve public health and security. Ultimately, because the battles ahead will no longer be just about health, this call for more open data sharing is a call to other sectors as well. We need to resolve our data issues posthaste to secure public trust in the plans, decisions, and pronouncements of the government and its private partners.

For questions or clarifications related to the technical or other aspects of this policy note, please send an email to upri.covid19@up.edu.ph. Scientific reports related to this statement will be posted in the endcov.ph site.