Puerto Rico
Science, Technology
& Research Trust

COVID-19 Data Management and Analysis Recommendations

A Report from the Data Science Advisory Committee, appointed by the Puerto Rico Science, Technology and Research Trust

November 10, 2020

Since the COVID-19 pandemic began, the Puerto Rico Science, Technology and Research Trust (the Trust) has been assisting the government of Puerto Rico in expanding its COVID-19 testing, data management, and community education capacity in an unofficial role, as part of its mission to facilitate and build capacity to continually advance Puerto Rico's economy and its citizens' well-being through science and technology. Most recently, this role has included participating in meetings to analyze available data related to the pandemic in Puerto Rico. To collect recommendations that could benefit the government in interpreting the available data related to the COVID-19 pandemic, the Trust's Scientific Committee of Trustees appointed a Data Science Advisory Committee with statistical, informatics, and epidemiological experts. Below, we present the analysis and recommendations of this group to be considered for improvement of health surveillance systems, both related to the pandemic and more broadly, and to assure the health, safety and rights to information of the Puerto Rican public.

Executive Summary

- The committee is greatly concerned with the way data has been managed, analyzed, reported, and interpreted during the pandemic. We are particularly concerned that continuing the current approach will exacerbate an already precarious situation.

- To assure timely, reliable, and accessible data that may be used by government officials and public health experts, Puerto Rico's Department of Health (PRDOH) should implement standards and operating procedures that prioritize electronic public health data collection. Data must be stored using a database management system (DBMS) rather than spreadsheets. The Bioportal DBMS is an excellent resource and should be the central repository for public health data including COVID-19 data.

- PRDOH should prioritize automating the operating procedures to keep the information in the DBMS as current as possible. As this is a National Emergency, the PRDOH has a vested interest in ensuring that there are data transmission solutions that can work seamlessly with the most common Laboratory Information Management Systems (LIMS) and Health Information Systems.

- Publicly accessible daily COVID-19 analyses and reports should be generated automatically using reproducible and publicly documented code. Microsoft Excel, Prism and other point-and-click tools that do not document the analysis and processing methodology of the data should not be used to prepare analyses for official reports.

- PRDOH should avoid over-summarizing COVID-19 related data into one number. Positivity rate, cases per capita, hospitalization, ICU usage, and deaths should be used separately to monitor the severity of the pandemic. This should be done at the global Puerto Rico-wide and local (town) levels.

- Data systems should be prepared to best support contact tracing efforts. We are encouraged by the formation of the *Sistema Municipal de Investigación de Casos y Rastreo de Contactos de Puerto Rico* (SMICRC).

- PRDOH should invest in statistical and informatics workforce and training, and use the lessons learned from COVID-19 and the Bioportal to implement similar changes across other health metrics, databases, and surveillance systems.

Introduction

It is critical for Puerto Rico to strengthen its public health data-related infrastructure, particularly as it relates to epidemiology, statistical analysis, and emergency response. Health outcomes data is the essential building block of all of this work. Rigorous data analysis of up-to-date and reliable data is key to effective decision making.

The World Health Organization (WHO) defines public health surveillance as “the continuous, systematic collection, analysis and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice” and calls it the “bedrock of outbreak and epidemic response.” Detecting and responding effectively to emergencies like COVID-19 requires timely and accurate data collection and management.

Disease surveillance in Puerto Rico is constrained by poor coordination of surveillance activities at the Department of Health (PRDOH). Many surveillance systems run as independent instruments, focused on internal data processes instead of being part of a comprehensive public health overview. For example, the surveillance systems of influenza, vector-borne diseases, and chronic illnesses work independently and are not part of an integrated system architecture. Furthermore, ad-hoc spreadsheets are used as master databases and many systems use manual data curation, which delays processes and may introduce human error. These hurdles have prevented the PRDOH Department from developing disease control and prevention strategies that are critical in response to pandemics like COVID-19. Multiple reporting systems, unclear lines of authority in the event of an outbreak, poor integration of private laboratories and stakeholders into public health systems, and lack of engagement and collaboration with private health care providers have combined to further hamper surveillance efforts in Puerto Rico. This was made evident during the aftermath of Hurricane María, during the Chikungunya epidemic, and now with the COVID-19 pandemic.

In early April, the PRDOH had few surveillance workers at the Department who possessed modern epidemiological, statistical, and programming skills necessary to rigorously manage and analyze large amounts of data from various sources. **We do note that this situation has improved during the last months, with the formation of the PRDOH Data Science team. However, the unclear lines of authority have made it difficult for this group to implement data science best practices across the PRDOH.**

In this report we provide specific recommendations related to the management and analysis of COVID-19 pandemic data. The recommendations are based on best practices as we have seen implemented in other jurisdictions, including Massachusetts.

Recommendations

Use of a Database Management System

We recommend that PRDOH use a database management system (DBMS) to collect and organize data. The current approach of defining a spreadsheet file as the *master* dataset should be discontinued. We understand that the Data Science team has implemented a DBMS, referred to here as the Bioportal. The Bioportal seems to meet all best practice standards and should become the official PRDOH COVID-19 database for all data related to COVID-19 testing and cases.

Use national standards for data transfers

All public and private reference and clinical laboratories, testing sites, and healthcare providers should be required to provide data to PRDOH in a timely and standardized manner. Cases, tests, hospitalizations, ICU usage, and deaths must be reported within 24 hours using standardized coding schemas. The national standard for reporting health data, Health Level Seven (HL7) should be required. To facilitate the transfer of large datasets generated by the hundreds of daily tests being conducted, we recommend the Fast Healthcare Interoperability (FHIR) resources be implemented. The PRDOH should impose penalties to healthcare organizations that do not comply with data entry and reporting requirements. PRDOH should also prioritize and allocate resources to helping these organizations become compliant.

We strongly recommend that **one standard** be implemented.

Report simple and standard metrics

The PRDOH has not been consistent or transparent with the analysis methodologies used when reporting or discussing various new COVID-19 metrics. In addition, we have seen the appearance of a number of COVID-19 scales with confusing use of colors and odometers, with undefined parameters and with unvalidated metrics. We recommend this be stopped. Instead simple and universally recognized standard metrics should be used and all metrics should be carefully defined to allow reproduction by others. Specifically, we recommend the following metrics be computed on a daily basis.

- **Daily 7-day average of positivity rate** - Positivity should be defined as the number of people who tested positive by DNA-based methods (i.e. molecular tests, PCR tests, LAMP-based tests) divided by the number of people who were tested by the same methods. Due to natural variability and strong weekend effects we recommend that positivity be calculated as running one-week average ending on the day in question. This metric, and in particular this definition is important because it provides information that cannot be gathered from just New Cases per Day or from Total Number of Tests. **Acceptable levels: The World Health Organization recommend 14 days below 5% before reopening the economy**. We are aware that there are other definitions of this rate which we do not recommend using. However, consistency and transparency are ultimately what is most important. Whichever definition of positivity is used, it should be clearly defined and uniformly used.

- **New cases per day –** This should be reported in a daily manner and a running 7-day average should be provided.

- **Daily number of COVID-19 patients that are hospitalized**.

- **ICU bed usage per day**: defined as the percent of ICU beds being used by COVID-19 patients, among beds not used by patients. **Acceptable level: Below 50%.**

- **COIVD-19 deaths per day**.

We do not see a need to calculate the reproduction number (R_T) as it does not provide any information that can't be inferred from visually inspecting the cases per day. Furthermore, increased or decreased testing can lead to an artificial reduction or increase in R_T, respectively, that may be misinterpreted.

Because each metric can fail in different ways, we recommend against combining them into one summary. Instead we recommend that graphical summaries for each of these metrics be carefully examined daily and shared with the public. They should be monitored for concerning trends such as multiple days with observed increases. Positivity and cases metrics should be reported at the Puerto Rico-wide and local (town) level. Map visualizations should be used to look for potential geographical outbreaks.

The *Sistema Municipal de Investigación de Casos y Rastreo de Contactos de Puerto Rico* (SMICRC), as well as San Juan's contact tracing system and the PRDOH state-wide, district-wide contact tracing systems should be consulted with respect to which metrics from those systems should also be reported, how, and when.

Assure all calculations are reproducible

By *reproducible calculations* we mean that the code and data is made available to reproduce the calculations without human intervention. **We strongly recommend against the use of point and click programs, such Microsoft Excel and Prism,** which do not document the analysis and processing methodology of the data. Recommended tools are R, Python, and SAS, with a strong preference for free and open source ones such as R and Python.

The daily report should be automatically produced directly from the database, without human intervention. We do recommend a human check for outliers or obvious errors of the report before it is made public. This check should take no more than 3 hours. If mistakes are found, they should be fixed in the database, and the report regenerated with the automated process. We recommend against editing the reports by hand as this results in un-reproducible procedures.

To provide an example that underscores the importance of reproducible calculations, we note that in a document prepared by Puerto Rico Public Health Trust's Mathematical and Epidemiological Modeling Team on August 25, 2020, and presented to the PRDOH on or before September 10, 2020, the positivity rate was shown to be 3.8% and the risk level claimed to be *medium*, a 4 in a scale of 0 (lowest) to 12 (highest). On September 10, 2020 the governor of Puerto Rico announced a new executive order opening gyms, movie theaters, and casinos. The calculations made in that document do not match the data shared by the PRDOH through the Bioportal database. Furthermore, we have not been unable to obtain code to reproduce these numbers and can only assume there are errors in the calculations. We note that in September, Puerto Rico saw a record 236 COVID-19 related deaths.

Invest in data science, programming, and statistics workforce and training

PRDOH needs to invest in the hiring and training of public health professions to develop key statistical and informatics skills with a goal of streamlining surveillance processes. Training in coding with data seems particularly important. The PRDOH Data Science team is a great. This group should be expanded and should be enabled to work closely with all PRDOH departments and offices that require updating of data systems. Once COVID-19 systems have been updated and the recommendations in this document implemented, the Data Science team would be in a good position to begin working with other health surveillance systems according to the priorities of the PRDOH.

Conclusions

In this report we provide specific recommendations related to COVID-19. In particular, we highlighted the importance of using a DBMS, automatized and reproducible calculations, and simple, transparent and interpretable metrics. However, the recommendations given here do not apply only to the current situation. Puerto Rico needs to transform and modernize the Island's surveillance systems, demonstrate rapid improvements, and inspire trust with surveillance partners in the field and the general public. Efforts should be made to consolidate its public health surveillance and information infrastructure into an integrated system and support data exchange according to national standards. While a wide range of diverse individual information systems can continue to exist, these systems must be coordinated, interconnected, comparable, and easy to use. Establishing standard operating procedures for core surveillance and response activities, building the Department's capacity for rapid response teams, and linking outbreak response structures with broader emergency management arrangements will enhance the utility of the surveillance system. Modernizing the Island's public health surveillance systems should be a top priority for Puerto Rico. The formation of the Data Science Team is a good start but much more is needed.

Prepared by the Data Science Advisory Committee Puerto Rico Science and Technology Trust

Members:

Rafael A. Irizarry (chair)
Professor and Chair, Department of Data Science, Dana-Farber University
Professor, Department of Biostatistics, Harvard T.H. Chan School of Public Health

María Eglée Pérez Hernandez
Professor and Chair, Department of Mathematics, University of Puerto Rico, Río Piedras

Ad-hoc members:

Gillian Haney
Director of Office of Integrated Surveillance and Informatics Services,
Bureau of Infectious Disease and Laboratory Sciences, Massachusetts Department of Public Health

Marc Lipstich
Professor, Department of Epidemiology, Harvard T.H. Chan School of Public Health
Director, Center for Communicable Disease Dynamics