What the COVID-19 Pandemic Has Reinforced: The Need for Accurate Data

Simone Arvisais-Anhalt, MD[1]

Christoph U. Lehmann, MD[2,3,4,5]

Jason Y. Park, MD, PhD[1,6]

Ellen Araj, MD[1]

Michael Holcomb, MS[4]

Andrew R. Jamieson, PhD[4]

Samuel McDonald, MD, MS[2,7]

Richard J Medford, MD[8, 2]

Trish M. Perl, MD, MSc[8]

Seth M. Toomay, MD[9]

Amy E. Hughes, PhD[5]

Melissa L McPheeters, PhD, MPH[10,11,12]

Mujeeb Basit, MD[2,13]

1. Department of Pathology, University of Texas Southwestern Medical Center, Dallas, Texas, USA
2. Clinical Informatics Center, University of Texas Southwestern Medical Center, Dallas, Texas, USA
3. Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, Texas, USA
4. Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, Texas, USA
5. Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, USA
6. Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, Texas, USA
7. Department of Emergency Medicine, University of Texas Southwestern Medical Center, Dallas, Texas, USA
8. Department of Internal Medicine, Division of Infectious Diseases and Geographic Medicine, University of Texas Southwestern Medical Center, Dallas, Texas, USA
9. Department of Radiology, University of Texas Southwestern Medical Center, Dallas, Texas, USA
10. Department of Health Policy, Vanderbilt University, Nashville, Tennessee, USA
11. Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA
12. Center for Improving the Public's Health through Informatics, Vanderbilt University
13. Department of Internal Medicine, Division of Cardiology, University of Texas Southwestern Medical Center, Dallas, Texas, USA

Corresponding Author:

Simone Arvisais-Anhalt, MD

Email: SIMONE.ARVISAIS-ANHALT@phhs.org

Department of Pathology, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas. United States.

Tel. 214.648.4088

Abstract:

The COVID-19 pandemic has challenged the United States' existing national public health informatics infrastructure. This report details the factors that have contributed to COVID-19 data inaccuracies and reporting delays and their effect on the modeling and monitoring of the COVID-19 pandemic.

Keywords: COVID-19, data modeling, public health informatics, information technology infrastructure, data integrity

The COVID-19 pandemic has challenged the United States' existing national public health and laboratory infrastructure. Central to fighting the pandemic is accurate and timely reporting of COVID-19 tests and case patients. Achieving this requires multiple elements including: unambiguous data requests; clearly defined variables; consistent labeling and processing of laboratory samples; uniform reports; accurate and complete data collection, aggregation, and transmission; and trained personnel to identify quality issues. In the US, COVID-19 interventions and prevention strategies have been decentralized and not harmonized across jurisdictions. Data complexity has resulted in reporting inaccuracies and delays at all levels of government, curtailing efforts to interrupt the spread of COVID-19.

COVID-19 Data Requests

Multiple governmental and non-governmental entities' requests for COVID-19 related data from healthcare institutions are duplicative, ambiguous, underspecified and lacking in granularity. Despite efforts to limit information gathering to the extent sufficient for informing policy and the public, the various requests commonly overlap, revealing poor communication between data requestors and lack of understanding of data availability, structure, and organization.

Entities with legitimate data needs commonly require reports in various formats (faxes, pdfs, csv or xls files, emails, HL7 feeds, online portals), and have varying deadline intervals (daily, weekly, monthly). Currently, our regional health care system reports to seven organizations (2 City, 1 County, 1 Regional, 1 State, 2 National). Data requests include information about testing (6/7), hospital census (5/7), ventilators (3/7), and staffing/capacity (1/7). Although electronic case reporting from established electronic

health records (EHRs) to public health agencies can be automated, uniform adoption has not been achieved due to high costs and complexity to vendors and institutions.

The uncoordinated data collection, lack of sharing among entities, and inability of the EHR to automatically generate report required us to create a reporting team composed of 11.85 full-time equivalents from the clinical laboratory, data warehouse, hospital operations, and hospital quality team. Not every institution has the required resources to create these complex reports.

Infectious Diseases Reporting Requirements

Data requests from public health and other authorities should be precise, actionable, and unambiguous. Early in the COVID-19 pandemic, confusion existed as to which date should be attributed to a positive test result (e.g., date of first symptoms, specimen collection, specimen resulting, or specimen reporting).[1] For a single specimen, these dates varied by weeks.

Reporting requests lacking necessary specificity have led to confusion. The White House Coronavirus Taskforce requested daily COVID-19 test results from hospital-based laboratories.[2] The instructions (since reversed due to duplicate reporting) were "*If all of your COVID-19 testing is sent out to private labs and performed by one of the commercial laboratories on the list below, you do not need to report using this spreadsheet.*" Hence, patient samples collected at institutions not performing COVID-19 testing and sent to one of the specified commercial laboratories for processing resulted in ONE report to the government. However, patient samples collected at institutions not performing COVID-19 testing and sent to laboratories for processing that were not on the

list may have been reported TWICE to the government – once by the hospital laboratory sending the specimen and once by the laboratory processing the specimen.

Ambiguity was also seen in local (city, county, and state) data requests. Researchers voiced concerns about COVID-19 results being reported multiple times[3] and some requesting entities revealed that they could not address duplicated results. For example, the Texas Department of State Health Services is unable to de-duplicate the test numbers from private laboratories.[4] Because the extent and detection of these errors across entities is unknown, data quality remains in doubt.

Missing Data Burden

Fulfilling data requests by different entities is challenging for laboratories when associated specimen data are missing. Accredited laboratories require two forms of patient identifiers (e.g., name and date of birth), collection date, and collection time to accept specimens. While SARS-CoV-2 testing sites often provide the minimum data to meet laboratory acceptance criteria, they do not uniformly collect all data requested by the various entities. Specimens are frequently shipped from third parties (e.g. nursing homes, physician offices, etc.) to affiliated hospitals to be shipped to reference laboratories. When receiving specimens with missing reportable information, the laboratory must either reject the specimen, delay specimen processing until information is received, or process the specimen with missing reportable information. In the latter case, the laboratory must choose to delay reporting or report incomplete information. For our medical center, collecting missing reportable information is a manual, unfunded process that has taken between one and 20 days. Once missing data are collected and reported, downstream practices can affect data accuracy depending on entities' reporting policies. Using the date a result was reported to

an entity versus specimen collection date will skew the shape and the trajectory of the epidemic curve.

Local Health Jurisdictions

Local and regional data aggregation strategies contribute to reporting delays depending upon the automation, integration, and interoperability of public health departments' electronic systems. Public health information technology infrastructure and capabilities vary widely. Resources allocated to local health departments depend on the investment and infrastructure provided by the federal government, state health departments, and categorical funding through initiatives and grants. Because categorical funding addresses specific public health goals, this funding is rarely used to develop or maintain flexible infrastructures to address future issues. Prior to COVID-19, health departments across the country received reportable infectious disease information through a combination of electronic laboratory reporting systems, faxes, and emails.[5] Health department staffing was usually sufficient to handle the volume of disparate reports. The volume generated by SARS-CoV-2 testing increased demands and stressed resources creating reporting failures.[6]

COVID-19 Disease Monitoring and Forecasting

The confluence of the aforementioned problems, in addition to reporting timeframes at an institutional level (Figure 1), resulted in inaccurate monitoring of the pandemic. For example, in Dallas County from April 27, 2020 to May 27, 2020, the daily incidence of COVID-19 positive cases artifactually decreased according to data reported by authorities on June 2 (Figure 2a). However, on June 9 after additional backlogged data were added for same timeframe, the revised data show the daily incidence had been stable or increasing.

The two different daily incidences for the same time period affected the calculated Rt, the virus's transmission rate (Figure 2b). Using the data from June 2, the Rt for May 27 was 0.73 - a rate that could quench the spread of disease. However, using the revised data from June 9, the Rt for May 27 was 1.03 - a rate that could fuel the spread of disease. The data discrepancy affected SIR (Susceptible, Infectious, or Recovered) modeling for the estimated and 14-day projection of COVID-19 cases (Figure 2c). For the June 2 data, the projected COVID-19 case numbers in Dallas County were 1,005; however, when corrected with backlogged data reported on June 9, the projected COVID-19 cases were 2,054. Data reporting delays directly affected policymakers' ability to interpret the trajectory of the epidemic, evaluate trends in cases and SARS-CoV-2 testing penetration and capacity, assess the role of interventions, and respond to the emerging data in a timely manner.

Conclusions

To successfully address an emerging pathogenic infectious disease with pandemic potential and limited public health and therapeutic interventions, like COVID-19 or H1N1, the availability of accurate, reliable, and timely laboratory and diagnostic data is critical for researchers, policy makers, and the public. Informatics pitfalls experienced during the H1N1 pandemic,[7] now re-experienced during the COVID-19 pandemic have shown how data robustness complements surveillance strategies and assures that public health and government entities can identify and trace cases and contacts, assess disease trends and patterns, and evaluate the epidemiology and transmission dynamics and assess mitigation strategies.

It is imperative that our national informatics strategy be improved upon. Local, regional, and national partners who share common data needs need to identify a finite list of clearly

defined variables for a minimum reportable data set to share across entities. This will reduce the reporting burden felt by healthcare providers.[6, 8] Existing public health information technology infrastructure must assure high-quality data and enable real-time data sharing. Healthcare institutions cannot opt-out of electronic reporting to public health agencies. Electronic health record vendors must increase interoperability among institutions and facilitate the basic public health reporting requirements to local, regional, and national public health entities without additional cost to an institution. Unique identifiers, like a universal healthcare identifier, should be implemented to minimize errors in providing needed demographic information and increase the integration of healthcare data with non-healthcare data. The restrictions that limit data sharing through the Health Insurance Portability and Accountability Act (HIPAA) should be revisited and criteria should be developed for the use of novel digital data, such a cellular data, with transparent public communication.[3, 9-11] Increased data management practices must be developed to avoid preventable methodologic and interpretation errors, such as by standardizing analytic strategies [12, 13] and publishing best practices for data presentation through dashboards.[14, 15] These investments will facilitate responses to the COVID-19 pandemic and to the inevitable next infectious disease outbreak.

Notes:

The authors of this article have no conflicts of interest to disclose.

References:

1. Landman K. Why is coronavirus data so damn difficult to communicate? Errors, lag, and perplexing charts—trying to understand COVID-19 data has become a major headache for many Georgians. Here's why. Available at: https://www.atlantamagazine.com/news-culture-articles/why-is-coronavirus-data-so-damn-difficult-to-communicate/. Accessed July 01, 2020.
2. Text of a Letter from the Vice President to Hospital Administrators. Available at: https://www.whitehouse.gov/briefings-statements/text-letter-vice-president-hospital-administrators/. Accessed July 02, 2020.
3. O'Reilly-Shah VN, Gentry KR, Van Cleve W, Kendale SM, Jabaley CS, Long DR. The COVID-19 pandemic highlights shortcomings in U.S. healthcare informatics infrastructure: a call to action. Anesth Analg **2020**.
4. Texas Case Counts COVID-19 Coronavirus Disease 2019. Available at: https://txdshs.maps.arcgis.com/apps/opsdashboard/index.html#/0d8bdf9be927459d9cb11b9eaef6101f. Accessed July 01, 2020.
5. Kilff S, Sanger-Katz M. Bottleneck for U.S. Coronavirus Response: The Fax Machine Before public health officials can manage the pandemic, they must deal with a broken data system that sends incomplete results in formats they can't easily use. Available at: https://www.nytimes.com/2020/07/13/upshot/coronavirus-response-fax-machines.html. Accessed September 19, 2020.
6. Foraker RE, Lai AM, Kannampallil TG, Woeltje KF, Trolard AM, Payne PRO. Transmission dynamics: Data sharing in theCOVID-19 era. Learn Health Syst: 8.
7. Infectious Disease Infrastructure: Impact and Continued Improvements Due to H1N1 Investments. Available at: https://www.astho.org/Infectious-Disease/Infectious-Disease-Infrastructure-Impact-and-Continued-Improvements-Due-to-H1N1-Investments/. Accessed October 23, 2020.

8. Kannampallil TG, Foraker RE, Lai AM, Woeltje KF, Payne PRO. When past is not a prologue: Adapting informatics practice during a pandemic. Journal of the American Medical Informatics Association **2020**; 27(7): 1142-6.
9. Lenert L, McSwain BY. Balancing health privacy, health information exchange, and research in the context of the COVID-19 pandemic. Journal of the American Medical Informatics Association **2020**; 27(6): 963-6.
10. Ienca M, Vayena E. On the responsible use of digital data to tackle the COVID-19 pandemic. Nature Medicine **2020**; 26(4): 463-4.
11. Sittig DF, Singh H. COVID-19 and the Need for a National Health Information Technology Infrastructure. JAMA **2020**; 323(23):2373-2374
12. Pearce N, Vandenbroucke JP, VanderWeele TJ, Greenland S. Accurate Statistics on COVID-19 Are Essential for Policy Guidance and Decisions. American Journal of Public Health **2020**; 110(7): 949-51.
13. Wolkewitz M, Puljak L. Methodological challenges of analysing COVID-19 data during the pandemic. BMC Med Res Methodol **2020**; 20(1): 81.
14. Zylla E, Hartman L. State COVID-19 Data Dashboards. Available at: https://www.shvs.org/state-covid-19-data-dashboards/. Accessed August 08, 2020.
15. Tracking COVID-19 in the United States From Information Catastrophe to Empowered Communities. Available at: https://preventepidemics.org/wp-content/uploads/2020/07/RTSL_Tracking-COVID-19-in-the-United-States_-7-23-2020.pdf. Accessed August 08, 2020.

Figure 1. Median and range of time (with possible sources of delay listed) for the stages from COVID-19 symptoms onset to data publication by requesting entity between 4/27/2020 and 5/27/2020 at a large academic medical center which also serves as a reference laboratory. In 689 of 12,182 (5.7%) SARS-CoV-2 tests performed, patients were asked about date of symptom onset at the point of specimen collection with a median of 3 days (0 – 31 days) of symptoms onset prior to specimen collection. The median turnaround time from specimen collection to specimen result for all 12,182 specimens was 32 hours (< 1 – 151 hours). The median courier time to the laboratory for all 12,182 specimens was 10 hours (<1 – 92 hours). The median turnaround time for the laboratory to report to a requesting entity for all 12,182 specimens was 29 hours (2-157 hours). It is unknown how much time requesting entities needed to publish the data they received.

Figure 2. a) Daily incidence of COVID-19 positive cases reported by the Dallas County Health Department based on specimen collection date from April 27 to May 27. Grey bars represent the daily incidence for the timeframe as reported on June 2. Black bars represent additional cases (backlog) for the same timeframe revised on June 9, 2020. b) Estimated COVID-19 R_t values from April 27 to May 27 using the incidence of positive cases reported on June 2 (grey line) and the revised R_t based on additional cases revised on June 9 (black line). c) SIR modeling for the estimated and 14-day projected trajectory of COVID-19 positive cases using the incidence of positive cases reported on June 2 (grey solid line estimated, grey dashed line projected) and when backlogged cases were added on June 9 (black solid line estimated, black dashed line projected).

Figure 1

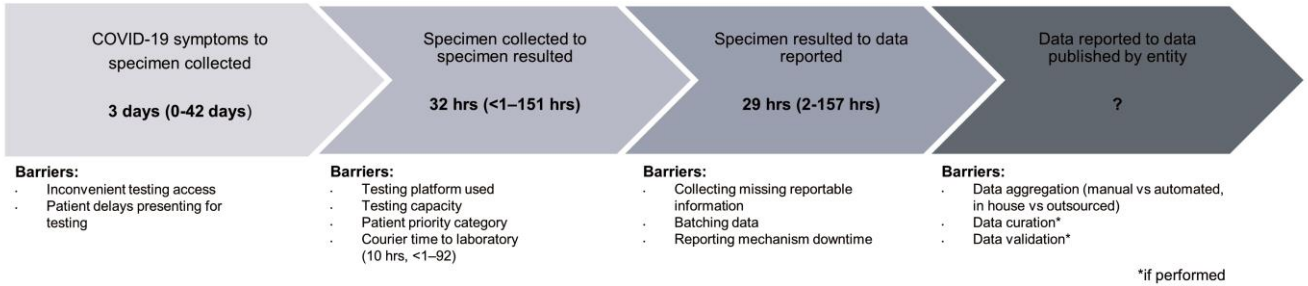
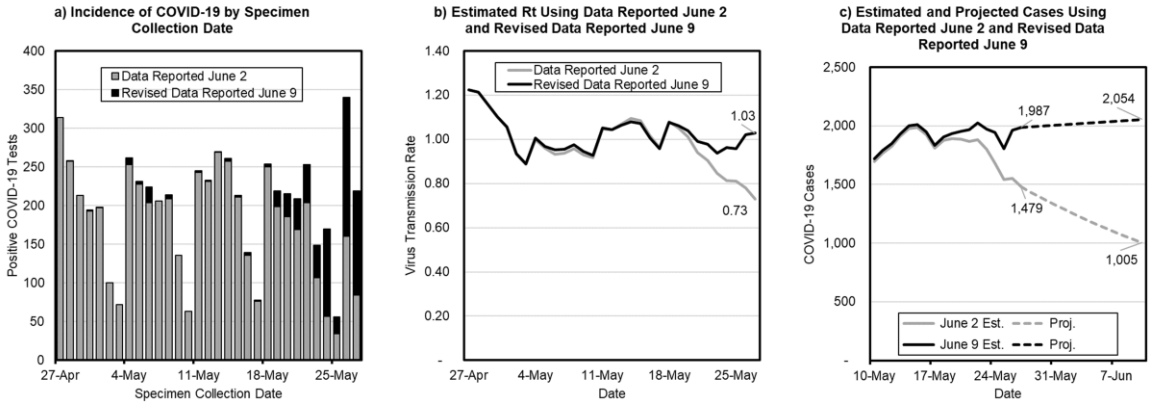

COVID-19 symptoms to specimen collected
3 days (0-42 days)
Specimen collected to specimen resulted
32 hrs (<1–151 hrs)
Specimen resulted to data reported
29 hrs (2-157 hrs)
Data reported to data published by entity
?
Barriers:
Inconvenient testing access
Patient delays presenting for testing
Barriers:
Testing platform used
Testing capacity
Patient priority category
Courier time to laboratory (10 hrs, <1–92)
Barriers:
Collecting missing reportable information
Batching data
Reporting mechanism downtime
Barriers:
Data aggregation (manual vs automated, in house vs outsourced)
Data curation*
Data validation*
*if performed

Figure 2


a) Incidence of COVID-19 by Specimen Collection Date
Positive COVID-19 Tests
Data Reported June 2
Revised Data Reported June 9
400
350
300
250
200
150
100
50
0
27-Apr
4-May
11-May
18-May
25-May
Specimen Collection Date
b) Estimated Rt Using Data Reported June 2 and Revised Data Reported June 9
Virus Transmission Rate
Data Reported June 2
Revised Data Reported June 9
1.40
1.20
1.00
0.80
0.60
0.40
0.20
-
1.03
0.73
27-Apr
4-May
11-May
18-May
25-May
Date
c) Estimated and Projected Cases Using Data Reported June 2 and Revised Data Reported June 9
COVID-19 Cases
2,500
2,000
1,500
1,000
500
-
1,987
2,054
1,479
1,005
June 2 Est.
Proj.
June 9 Est.
Proj.
10-May
17-May
24-May
31-May
7-Jun
Date