Infection Control & Hospital Epidemiology (2023), **44**, 802–804
doi:10.1017/ice.2022.55

Concise Communication

Hospital coronavirus disease 2019 (COVID-19) public health reporting: Lessons from validation of an automated surveillance tool to facilitate data collection

Shantini D. Gamage PhD, MPH[1,2], Martin E. Evans MD[1,3,4], Brian P. McCauley DPM[1], Karen R. Lipscomb MSN[1], Linda Flarida MBA[1], Makoto M. Jones MD[5,6], Michael Baza BS[5,6], Jeremy Barraza BS[5,6], Loretta A. Simbartl MS[1] and Gary A. Roselle MD[1,2,7]

[1]National Infectious Diseases Service, Specialty Care Services, Veterans Health Administration, US Department of Veterans Affairs, Washington, DC, [2]Division of Infectious Diseases, Department of Internal Medicine, University of Cincinnati College of Medicine, Cincinnati, Ohio, [3]Lexington Veterans Affairs (VA) Healthcare System, Lexington, Kentucky, [4]Division of Infectious Diseases, Department of Internal Medicine, University of Kentucky School of Medicine, Lexington, Kentucky, [5]VA Salt Lake City Healthcare System, Salt Lake City, Utah, [6]Divison of Epidemiology, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, Utah and [7]Cincinnati VA Medical Center, Cincinnati, Ohio

Abstract

A comparison of computer-extracted and facility-reported counts of hospitalized coronavirus disease 2019 (COVID-19) patients for public health reporting at 36 hospitals revealed 42% of days with matching counts between the data sources. Miscategorization of suspect cases was a primary driver of discordance. Clear reporting definitions and data validation facilitate emerging disease surveillance.

(Received 2 July 2021; accepted 21 February 2022; electronically published 30 March 2022)

The coronavirus disease (COVID-19) pandemic has had a significant impact on patient health and logistics in healthcare systems globally.[1,2] The Veterans' Health Administration (VHA) developed a surveillance system using electronic health record (EHR) data to monitor the impact of COVID-19 at the 170 VHA medical centers across the United States in real time.[3] The VHA leveraged this new system to report daily data for each VHA medical center to the Centers for Disease Control and Prevention (CDC) National Healthcare Safety Network (NHSN) COVID-19 Acute-Care Module.[4]

Although EHR data has been used for other VHA surveillance reporting,[5] application of centralized EHR data for emerging disease reporting can be challenging. Here, we compared the COVID-19 data extracted centrally to data collected by individual VHA facilities to validate the EHR extractions. Our findings provide insight into data collection for public health reporting of novel diseases.

Methods

Data collection was conducted for 10 business days in June 2020 to compare the daily computer-extracted data from the COVID-19 surveillance system (ie, computer-extracted data) being sent to NHSN with data manually collected from volunteer VHA medical centers (ie, facility-reported data) for the same period and using the same NHSN definition for the count of hospitalized patients with suspected and confirmed COVID-19 (see the Supplementary Information online for definitions and details on data collection and analysis). Total counts of hospitalized patients determined from computer extraction and facility reporting were compared by *t* test (SAS version 9.4 software, SAS Institute, Cary, NC).

For both data sources, personally identifiable information (ie, patient first name, last name, birthdate, and last 4 digits of the Social Security number) corresponding to the daily patients counted was collected to compare accuracy at the patient level (ie, even when counts match, the patients reported can be different). Additionally, for facilities that had zero or only 1 day of total counts matching between the 2 sources, the medical charts of patients identified by both sources were reviewed to determine whether they should have been included in the count based on the NHSN definition.

This work was approved by VHA Central Office as a validation effort for operational systems improvement.

Results

Of 170 VHA facilities, 36 (21%) volunteered to participate in the project, resulting in 356 days of facility data (Table 1 and Fig. 1; see Supplementary Table S1 online for facility characteristics). Overall, the study included 1,472 patient days for computer-extracted data and 1,353 patient days for facility-reported data ($P = .34$) (Table 1). The count of hospitalized patients was the same for the 2 data sources in 151 (42%) of the 356 facility days reported. For 139

Author for correspondence: Shantini D. Gamage, National Infectious Diseases Service, Veterans Health Administration, Department of Veterans Affairs. E-mail: shantini.gamage@va.gov

Cite this article: Gamage SD, *et al.* (2023). Hospital coronavirus disease 2019 (COVID-19) public health reporting: Lessons from validation of an automated surveillance tool to facilitate data collection. *Infection Control & Hospital Epidemiology*, 44: 802–804, https://doi.org/10.1017/ice.2022.55

Table 1. Summary Statistics for All Participating Facilities

Characteristic	Total
Total no. of days for all facilities	356[a]
Hospitalized patients for all facility days, no.	
Computer-extracted[b]	1,472
Facility-reported[c]	1,353
Unique hospitalized patients for all facility days, no.	
Computer-extracted	559
Facility-reported	474
Unique patients from computer-extracted and facility-reported	813
Comparison of computer-extracted and facility-reported hospitalized patient counts each facility day, no. (%)	
Days counts matched	151/356 (42.4)
Count-matching days with 100% patient match[d]	139/151 (92.1)
Days counts matched or were discordant by 1 patient	261/356 (73.3)
Days counts matched or were discordant by up to 2 patients	301/356 (84.6)

Note. VHA, Veterans Health Administration.

[a] 34 (94%) of 36 facilities reported data for all 10 days of the project; 2 facilities reported data for a portion of the 10 days of the project (see Fig. 1 for more details).

[b] Data extracted by VHA automated data extraction system for daily reporting of VHA medical facility counts to the National Healthcare Safety Network.

[c] Data reported by VHA medical facilities participating in this review based on manual assessment of daily counts by facility staff.

[d] Patient matching at a facility was determined by comparison of patient identifiers (names, dates of birth, and last four digits in social security numbers) between the computer-extracted and facility-reported counts for each day.

(92.1%) of these 151 days, there were complete patient matches of personally identifiable information between the 2 sources (Table 1 and Fig. 1). When counts from the 2 sources were compared allowing for discordance by 1 or 2 patients, the percentage of facility days included increased from 42% to 73% (261 of 356) and 85% (301 of 356), respectively (Table 1).

Daily counts were also assessed for each facility, and the review of personally identifiable information showed variability in the extent of patient matching (Supplementary Table S2 online). The difference in counts between the 2 sources on nonmatching days was low overall; the median difference in count was 1 for 23 (68%) of 34 facilities. On days when patient counts matched on the facility level, the personally identifiable information of patients also tended to match; the mean for all facilities was 95% ±16%. However, on days with nonmatching counts between data sources, the mean matching of personally identifiable information for all facilities was only 48% ±30% (Supplementary Table S2; see also Figure 1 for variability in matching personally identifiable information).

Overall, 12 (33%) facilities met criteria for patient chart reviews. For 8 facilities, chart reviews showed that the data source with higher daily counts overcounted cases (Supplementary Table S3 online). Reasons for facility-reported overcounting included reporting patients who had a past positive test result but no symptoms on admission, having a protocol designating certain patients as suspect without symptoms (eg, transfers from nursing homes), universal laboratory screening of all patients for COVID-19 regardless of symptoms or reason for admission, and errors in following instructions. Computer-extracted overcounting occurred with reporting patients who had a past positive result but no symptoms on admission, patients with no COVID-19 testing or a negative result, and patients without symptoms but who were tested based on local universal screening policy.

Discussion

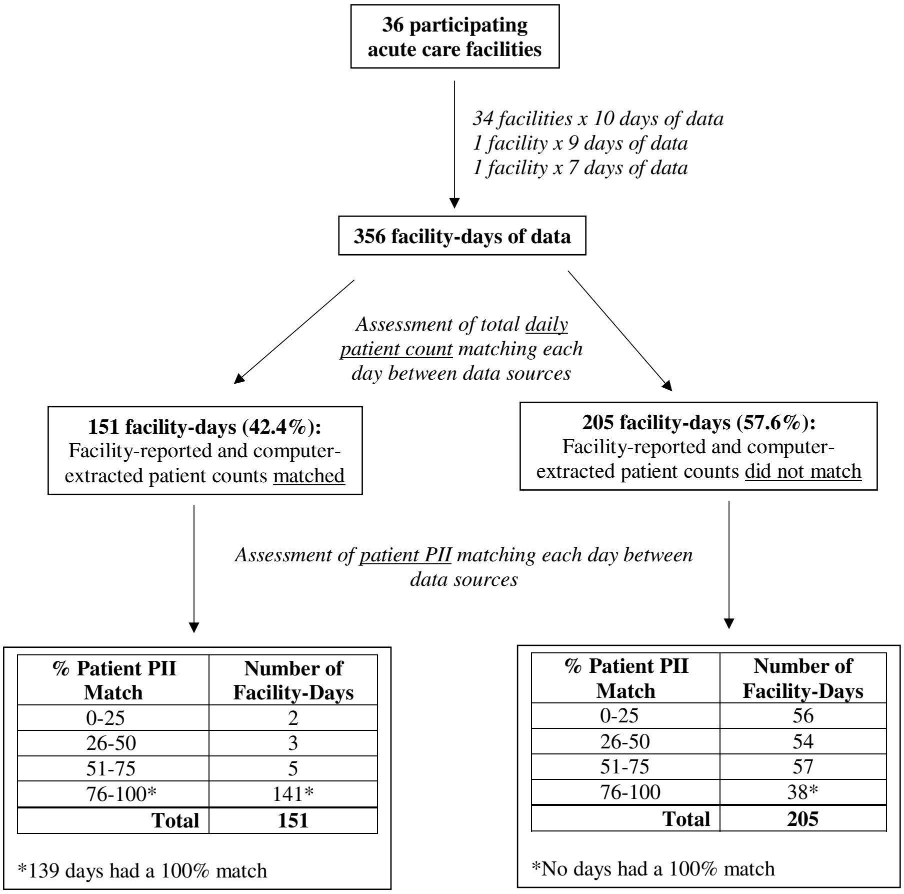
Electronic health record data are increasingly being used for public health surveillance.[6,7] Our findings offer insights for designing emerging disease surveillance systems, and they highlight the importance of periodic validation by those collecting data for reporting.

The hospitalized patient-count match between computer-extracted and facility-reported data was ~42%. For the days that counts did match, the individual patients from the 2 sources almost always matched, indicating a key level of accuracy. The most common reason for count mismatching in both sources, substantiated by comparing personally identifiable information and selected chart reviews, was determination of "suspect" cases for inclusion in the daily report though the impact varied by facility. Symptoms of COVID-19 are broad and nonspecific[8] and are especially difficult to discern in older people who may have atypical presentation and multiple comorbidities.[9,10] Therefore, identifying suspected cases by automated extraction was limited to those who had a specific tag in their electronic record such as being a "person under investigation." From a facility-reporting perspective, local criteria for suspected cases sometimes differed from the NHSN definition. These findings lead to several observations to improve emerging disease surveillance: (1) early standardization of healthcare system laboratory and health-record terms in the EHR for an emerging disease can facilitate data extraction; (2) capturing suspect cases is often necessary for emerging disease reporting systems, and clear definitions and instructions for suspect cases can prevent overreporting; and (3) having separate data elements for confirmed cases and suspect cases will lessen data interpretation issues.

Although we identified areas for improvement, computer extraction was useful for assessing the change in patient counts over time and often was more accurate than reporting by facilities. Allowing for a limited discrepancy in count matching between sources increased the percentage of days matching from 42% to 85%. And, when counts at facilities were discordant, the difference in counts between sources was usually low. In the early months of an emerging disease, leveraging EHRs for automated extraction can relieve the data-collection burden from staff with sufficient accuracy for monitoring changes in case counts.

This study had several limitations. Facility participation was voluntary, and several of the facilities had low hospitalized-patient counts. Our results may not reflect validation in high-incidence areas. Nonetheless, in an emerging biological event, even capturing small numbers of cases is critical, and validation in low-incidence areas is valuable. Another limitation is the unknown generalizability of the VHA extraction system for public health reporting in other healthcare systems; however, the reported challenges related to designation of suspect cases is informative for the development of any such system. Finally, we did not collect detailed information on how facilities were determining their counts; interpretations had to be drawn from selected patient chart reviews.

Surveillance during an emerging infectious disease depends on data collection definitions and reporting systems. Here, 2 data-collection mechanisms, computer extraction and manual facility reporting demonstrated comparable results for surveillance of disease incidence, with similar pitfalls related to the sometimes


36 participating acute care facilities
34 facilities x 10 days of data
1 facility x 9 days of data
1 facility x 7 days of data
356 facility-days of data
Assessment of total daily patient count matching each day between data sources
151 facility-days (42.4%): Facility-reported and computer-extracted patient counts matched
205 facility-days (57.6%): Facility-reported and computer-extracted patient counts did not match
Assessment of patient PII matching each day between data sources

% Patient PII Match	Number of Facility-Days
0-25	2
26-50	3
51-75	5
76-100*	141*
Total	**151**

*139 days had a 100% match

% Patient PII Match	Number of Facility-Days
0-25	56
26-50	54
51-75	57
76-100	38*
Total	**205**

*No days had a 100% match

Fig. 1. Summary of daily count and personally identifiable information (PII) matching between computer-extracted and facility-reported data for the 36 participating medical facilities. In the last row of boxes, each box shows the extent of personally identifiable information matching between the 2 data sources for each facility day, by quartiles.

ambiguous nature of case classifications. Validation studies are critical for identifying areas for improvement when developing data-collection platforms for emerging diseases.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/ice.2022.55

Acknowledgments. The authors are very grateful to the staff at the 36 participating facilities who collected data for this project. The views expressed are those of the authors and are not necessarily the official views of the Department of Veterans' Affairs.

Financial support. No external financial support was received for this work.

Conflicts of interest. All authors report no conflicts of interest relevant to this article.

References

1. Moghadas SM, Shoukat A, Fitzpatrick MC, *et al.* Projecting hospital utilization during the COVID-19 outbreaks in the United States. *Proc Nat Acad Sci U S A* 2020;117:9122–9126.
2. Crespo J, Fernanez Arrillo C, Iruzubieta P, *et al.* Massive impact of COVID-19 pandemic on gastroenterology and hepatology departments and doctors in Spain. *J Gastroenterol Hepatol* 2021;36:1627–1633.
3. COVID-19 national summary. Department of Veterans' Affairs website. https://www.accesstocare.va.gov/Healthcare/COVID19NationalSummary. Accessed October 28, 2021.
4. Sapiano MRP, Dudeck MA, Soe M, *et al.* Impact of coronavirus disease 2019 (COVID-19) on U.S. hospitals and patients. *Infect Control Hosp Epidemiol* 2022;43:32–39.
5. Jones BE, Haroldsen C, Madras-Kelly K, *et al.* In data we trust? Comparison of electronic versus manual abstraction of antimicrobial prescribing quality metrics for hospitalized veterans with pneumonia. *Med Care* 2018;56:626–633.
6. Salazar M, Stinson KE, Sillau SH, Good L, Newman LS. Web-based electronic health records improve data completeness and reduce medical discrepancies in employee vaccination programs. *Infect Control Hosp Epidemiol* 2012;33:84–86.
7. Klompas M, Murphy M, Lankiewicz J, *et al.* Harnessing electronic health records for public health surveillance. *Online J Public Health Inform* 2011;3:ojphi.v3i3.3794.
8. Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review. *JAMA* 2020;324:782–793.
9. Tay HS, Harwood R. Atypical presentation of COVID-19 in a frail older person. *Age Ageing* 2020;49:523–524.
10. Neumann-Podczaska A, Al-Saad SR, Karbowski LM, Chojnicki M, Tobis S, Wieczorowska-Tobis K. COVID-19 clinical picture in the elderly population: a qualitative systematic review. *Aging Dis* 2020;11:988–1008.