Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Commentary

Mind the clinical-analytic gap: Electronic health records and COVID-19 pandemic response

Sylvia E.K. Sudat [a,*], Sarah C. Robinson [a], Satish Mudiganti [a], Aravind Mani [b,c], Alice R. Pressman [a]

[a] Center for Health Systems Research, Division of Research, Development & Dissemination Sutter Health, Walnut Creek, CA, United States
[b] California Pacific Medical Center, San Francisco, CA, United States
[c] Pacific Inpatient Medical Group, San Francisco, CA, United States

ARTICLE INFO

ABSTRACT

Data quality is essential to the success of the most simple and the most complex analysis. In the context of the COVID-19 pandemic, large-scale data sharing across the US and around the world has played an important role in public health responses to the pandemic and has been crucial to understanding and predicting its likely course. In California, hospitals have been required to report a large volume of daily data related to COVID-19. In order to meet this need, electronic health records (EHRs) have played an important role, but the challenges of reporting high-quality data in real-time from EHR data sources have not been explored.

We describe some of the challenges of utilizing EHR data for this purpose from the perspective of a large, integrated, mixed-payer health system in northern California, US. We emphasize some of the inadequacies inherent to EHR data using several specific examples, and explore the clinical-analytic gap that forms the basis for some of these inadequacies. We highlight the need for data and analytics to be incorporated into the early stages of clinical crisis planning in order to utilize EHR data to full advantage. We further propose that lessons learned from the COVID-19 pandemic can result in the formation of collaborative teams joining clinical operations, informatics, data analytics, and research, ultimately resulting in improved data quality to support effective crisis response.

1. Introduction

Throughout the course of the COVID-19 pandemic, data sharing – across the United States (US) and between countries – has been extremely important, especially during early efforts to understand and predict the course of the crisis [1–3]. As in all data-based studies, these efforts will be directly impacted by data quality; poorly-defined features or flawed data sources can lead to irreproducible or even invalid results. This can have serious impacts, especially in the context of a pandemic, when we are looking to research and innovative analytics for insights and to guide our public health response.

While much focus is typically given to new analytic methods, conversations regarding data quality tend to be limited and under-emphasized, despite the existence of multiple conceptual frameworks for electronic health record (EHR) data quality assessment [4–13]. There have been some examples of informatics support or guidelines to leverage EHRs for rapid pandemic response [14–17], however

challenges with data quality during COVID-19 have been investigated mostly in the context of data sharing between countries [18–20]. Less attention has been given to the specific challenges faced by US health systems and hospitals in reporting real-time data to regulatory agencies, and the limitations of EHRs as sources of those data, even though these processes are the fundamental pipelines of information illuminating the progression of the pandemic.

There are multiple perspectives on what constitutes high-quality data, such as its accuracy as compared to some known gold standard, or its appropriateness for use in the context of a specific application (fitness for purpose) [11–13]. In the context of this paper, we use the term "data quality" to describe the ability of EHR-derived data to produce an accurate, reliable, and consistent aggregate-level picture of what is happening at the point-of-care. For example, a high-quality real-time data source should – under this definition – be able to produce a count of COVID-19-positive patients in a given hospital unit that agrees with the count obtained by asking the staff and clinicians directly caring

for the patients in that unit. This most closely aligns with the fitness-for-purpose definition of data quality.

As the US expands its system-based health care [21–24], reliable reporting using EHRs will become even more relevant, and lessons learned during this pandemic will be essential to promoting a thriving and robust clinical-analytic ecosystem that can utilize the EHR to quickly satisfy both operational and research information needs. As individual clinics and hospitals consolidate into health systems with shared EHRs, data analysts benefit from centralized patient information. This also leads, however, to more distance between the utilizers of data – analysts, researchers, data scientists – and the clinicians and staff who generate the data. Each clinic, hospital, and other care location can introduce variability when capturing patient information, which in turn can lead to misinterpretation of the resulting EHR data if open lines of communication between data creators and data consumers are not present. This effect is further magnified in a public health crisis, when data must also be aggregated across different health systems and care centers.

In this context, we describe challenges and lessons learned in responding to the COVID-19 daily reporting required by California hospitals from the perspective of Sutter Health, a large, integrated, mixed-payer health system in northern California, US. Sutter serves a patient base of approximately 3.5 million patients with 24 acute care hospitals, 5 medical foundations, a home care and hospice agency, and other health care services across 22 California counties. Specifically, we highlight the importance of the EHR in supporting pandemic response, and explore some of the barriers to producing high-quality, real-time data from an EHR data source. We propose that forming collaborative teams joining data and analytics to clinical operations in the early stages of health care crisis planning can help mitigate these challenges, resulting in a lower burden on clinical staff and higher-quality data.

2. The role of electronic health records in pandemic response

California hospitals were required to provide daily reports with metrics related to COVID-19 starting in March of 2020 [25,26]. At the time of this writing, the current version of the data dictionary includes 66 required data elements to be reported each day and 37 each week, encompassing domains including a snapshot of COVID-19 hospital occupancy, prior day information on COVID-related emergency department visits and inpatient admissions, general hospital occupancy, ventilator use, hospital capacity, use of surge beds, personal protective equipment resources, hospital staffing, and in-hospital deaths [25]. Additional domains related to influenza were added in late 2020. This is a large number of data elements, even for a single hospital; the number grows substantially for health systems that report out on multiple hospitals. For Sutter Health's 24 acute care hospitals, this means that 1584 total data elements are reported by noon each day with 888 additional data elements reported weekly. A small subset of these data, aggregated by county, is available publicly through the California Department of Public Health's (CDPH) open data portal [27].

The ability to quickly ascertain a high-level view of COVID-19-related hospital data is not only important for the state and the nation's pandemic response, it has been essential to guiding crisis planning within health systems. An ability to understand, for example, which intensive care units (ICUs) might allow for an influx of patients in the event that another ICU were to become overburdened, requires an ability to look across multiple hospitals and quickly obtain near real-time information about capacity and occupancy.

Ideally, EHRs should be well-positioned to respond to these urgent data and reporting needs. However, EHRs are created primarily for patient care and to facilitate billing, which may or may not ensure high-quality data in a readily usable format. When conducting research, typically time is taken to thoroughly review data sources, investigate any inconsistencies, and determine the best definitions for the various data elements that are needed. When collisions exist between

documentation practices and the need for high-quality research data-sets, technical workarounds can be employed to produce the needed data. For example, a programmer might dedicate weeks or months to validating a natural language processing (NLP) solution to extract a particular piece of important information from a physician's free-text notes. In addition, data elements required for reporting may simply not be available or be incomplete due to their irrelevance to patient care. As these data points become pertinent, task forces may work with clinicians and staff to increase the quality of the data capture and improve documentation procedures. Doing any of these types of data validation, processing, and cleaning is only possible, however, when the data do not have to be real-time.

During a pandemic, the data need to be both high quality and close to real-time. On a small scale, it may be possible to perform data validation on-the-fly, such as looking at charts to validate whether or not a patient actually has COVID-19, or whether a recent death was actually a COVID-19 death. However, when health systems are required to produce enterprise-wide reporting for multiple facilities, manual verification quickly becomes unworkable. The reporting task can become much less trivial when unmodified EHR data are being considered and reporting must be done for many facilities at once.

3. Data challenges

Table 1 presents some of the challenges we faced in sourcing data from our EHR for daily hospital reporting of COVID-19. These challenges have been grouped into data domains. While seemingly simple to identify within the EHR, each domain required extensive EHR exploration and validation.

3.1. Hospital capacity

We faced three major challenges in the hospital capacity domain. The first challenge was being able to accurately reproduce the number of acute beds in each hospital from the EHR bed records. Hospital bed counts are often reported based on licensing numbers, which may not reflect actual capacity on a given day; actual available bed counts can fluctuate up and down due to isolation requirements, room maintenance, staffing, etc. EHRs may also have electronic department structures that do not reflect the typical patient capacity – such as allocations of space to which patients could be assigned, for example, in a natural disaster when hospitals were overflowing and patients have to receive care in a non-standard area (like a hallway). For the Sutter system, the EHR's electronic version of the hospital's beds produced a count that differed substantially from counts expected based on licensure. This required manual review and adjudication to arrive at the appropriate number.

The second challenge arose with the need to classify beds by level of

Table 1
COVID data domains and challenges in reporting data from the EHR.

Data domain	Challenges
Hospital capacity	- Difficulty reproducing the actual bed count in the hospital from EHR bed records - Mismatch between bed classification according to use (e. g. telemetry) and capacity estimates based on location (e. g. department) - Changes in bed classifications and counts due to pandemic surge capacity coming online or offline
COVID-19 confirmed patients	- Difficulty identifying COVID-positive patients who had external (e.g. scanned) results - Inconsistency in fixed-field real-time diagnoses such as the hospital problem list
COVID-19 suspected patients	- Symptom-based definitions that are difficult to reproduce using EHR data
COVID deaths	- Difficulty identifying COVID-19 patient deaths with confidence without being able to utilize discharge diagnoses

care, such as ICU. For daily government agency reports, hospitals were required to provide a count of ICU beds and the number of occupied ICU beds, by COVID and non-COVID patients. In many hospitals, patients can receive an ICU level of care in a non-ICU unit, such as in a step-down unit, telemetry unit, or even a medical/surgical unit. These are ICU *patients*, but they are not occupying ICU *beds*. Conceivably, the number of ICU patients could therefore outstrip the number of dedicated ICU beds. This issue was magnified for telemetry beds (which were ultimately removed as a subcategory from the daily required report), because it is quite common for medical/surgical units to have at least some telemetry capability.

The third challenge arose in tracking surge capacity and determining current use of surge beds. We were fortunate in that our hospitals' surge beds were formally set up as "departments" in the EHR, which allowed us to see when the beds were activated or inactivated and to produce an accurate bed count. However, surge units were created within the EHR at various time points during the pandemic, and consistent communication was required to appropriately incorporate this information into the reporting structure. In addition, while hospitals had multiple strategies in place to accommodate surges, until departments were added to the EHR, electronic reporting could not account for this additional capacity.

3.2. Identification of COVID-19 patients and deaths

Three of the most important data elements related to COVID-19 hospital occupancy, COVID-confirmed, COVID-suspected, and in-hospital deaths, also posed difficulties for electronic reporting in real-time.

At the time of this writing, the CDPH defined COVID patients in the following ways:

- *COVID confirmed* – "an individual who has at least one respiratory specimen that tested positive for SARS-CoV2, the virus that causes COVID-19, by a CDC laboratory, state or local public health laboratory, or commercial laboratory using FDA-validated COVID-19 nucleic acid amplification (NAA)."
- *COVID suspected* – "persons without a laboratory-confirmed COVID diagnosis who, in accordance with CDC's Interim Public Health Guidance for Evaluating Persons Under Investigation (PUIs), have signs and symptoms compatible with COVID (most patients with confirmed COVID have developed fever and/or symptoms of acute respiratory illness, such as cough, shortness of breath or myalgia/fatigue)."

COVID *confirmed* is the easier to implement – patients who have a positive test result. Even that seemingly simple requirement, however, can become complex when extracting the data from an EHR. There are multiple ways that positive tests can be documented, and only some – those patients who were tested and resulted at labs within the health system or hospital – are easily extractable. External results may be stored as scanned documents or images, which makes easy electronic interpretation virtually impossible. This means that health systems may have to turn to other ways of identifying patients – such as in-hospital diagnoses, problem lists, or indications of COVID-19 isolation precautions. These data elements, however, may also be frequently present for COVID *suspected* cases who test negative, and thus not be reliable as a method of discerning COVID *confirmed* cases. COVID *suspected* is even more fraught with adversity. The definition is based largely on symptoms, which are very difficult to report from the EHR without mining text notes.

In-hospital COVID *deaths* also present difficulties for real-time reporting. In a retrospective study, a researcher might rely upon the hospital discharge diagnosis (assuming an International Classification of Diseases [ICD] diagnosis code for COVID-19 had been introduced) to identify in-hospital deaths among COVID patients, especially in

situations when a positive lab result may be difficult to find electronically. However, discharge diagnoses are not typically coded in the patient's chart until well after discharge, meaning that real-time reporting using this method is not possible. This complicates the identification of COVID *deaths*, especially when historical data cannot be updated to account for delayed identification.

The inadequacy of the EHR to meet the immediate reporting needs for these three key domains has several downstream effects. If immediate reporting is mandated, and cannot be consistently and reliably extracted from the EHR, then each hospital in a health system may have to take on manual reporting of required daily data. Not only does this add to the burden of those at the front lines of patient care, it also adds to the likelihood of variability in definitions and of data entry error. After hospitals have been required to take on manual reporting, clinicians and staff also may have diminished confidence in the ability of the EHR-based reporting to *ever* reliably obtain the required information.

4. A cultural divide

There is a profound separation between the generators of data – the clinicians and healthcare staff who document care in the EHR – and the consumers of data – the analysts and researchers who use the EHR data for decision support, research, and business intelligence. The documentation requirements at the point of care place a significant burden on clinicians and staff, and EHRs are often seen as impediments to their primary jobs of caring for patients. This can result in a cultural divide between the data consumers and the data generators. Analysts and researchers may feel frustrated by inconsistent documentation or free-text documentation in notes fields where a fixed field is available; clinicians and staff may feel that data consumers essentially do not understand patient care or how their facility operates, and feel resentful that they are being asked to change the way they are documenting when they are already overburdened with documentation requirements. This reflects an essential difference in goals: healthcare workers are appropriately motivated by providing the best care for patients, which may not always align with the goal of generating optimal datasets for research and analytics. A certain documentation practice – such as noting in a free-text field instead of in a fixed field – may pose no problems at all in the patient care setting, and in fact could make it easier for another care provider to gain an overall impression by simply glancing through the note. However, free-text note documentation can cause significant problems in the data analytics setting, making certain important data elements difficult or sometimes impossible to access depending upon the consistency of documentation and the resources available to the analytics teams.

5. Data and analytics: An essential part of the healthcare crisis response team

One of the main results of the divide that typically exists between research/data/analytics and clinical operations is that data and analytics personnel are not usually considered as part of the core team when clinical crisis plans are formed. This became evident during the COVID-19 pandemic. In our health system, we were eventually able to create a robust reporting infrastructure to support our hospitals and leverage our EHR. This was accomplished by assembling a collaborative, cross-functional team – including leaders in research, business intelligence, clinical informatics, and clinical quality. Only after this team was in place did it become possible to determine how best to define various high-impact metrics, and to work more closely with hospital staff as discrepancies arose. Individuals on the data side were able to gain insights into clinical workflows, and clinical leaders were able to better understand how different documentation practices might affect our health system's ability to report on various COVID-related measures. This does not mean that the challenges related to reporting real-time data out of the EHR ceased to exist. But we were able to work

together to determine the best ways to address data and definition challenges, and to increase the clinical staff's trust in the accuracy and reliability of centralized reporting. However, countless hours could have been saved – and likely higher-quality data produced earlier on – if this collaboration had been constructed in the early stages of crisis planning, instead of months later.

Other health systems may have faced similar challenges, or may have recognized earlier that cross-functional teams would be an important part of pandemic readiness [17]. When using aggregate-level data to help understand the course of this pandemic, we simply do not have the information necessary to understand how those data were generated, or how reporting practices and data quality may have changed over time. The opportunity to work toward better EHR-based reporting therefore remains broadly relevant.

6. Lessons learned

One of the results of the COVID-19 pandemic's urgency for health systems is that many collaborative relationships were established between previously unconnected departments and individuals. This may, therefore, be the ideal time to encourage a cultural shift in how we think about and engage with health data. As researchers, analysts, and data scientists, this means engaging more closely with clinical operations, and increasing our understanding of how EHR data are generated, not only how they are extracted. For those in clinical operations, this means engaging with researchers, analysts, and data scientists not only as service providers, but as part of the clinical team. As our society moves to a model of larger health systems, reliability and quality in EHR data will only become more important.

In a situation like the COVID-19 pandemic, we in health systems need to be able to produce high-quality data quickly, accurately, and consistently. However, our current EHR and informatics infrastructure may be ill-equipped to produce enterprise-level, high-quality reporting in real time. This can not only place an undue burden on those at the patient-care level to manually report required data elements, but inconsistent data quality has repercussions on our ability to make predictions and draw conclusions from those data.

We are unlikely – at least in the short term – to re-invent the structure of our EHRs. We can reinvent, however, the structure of our teams. There is an important opportunity to recognize the gaps in collaboration that exist, and to respond by working to close those gaps. This is a cultural change that can begin now, one which will support not only current efforts, but provide a collaborative framework for longer-term solutions. This framework can be used to work toward the goal of making future data more reliable and easier to produce – to fully unleash the power of our EHRs, and make sure that they can give us the insights we need to utilize cutting-edge methodologies to full advantage in responding to future public health crises.

7. Limitations

We recognize that we are presenting only a single perspective from one health system in California, US, which is likely not completely generalizable to all hospitals across the US or around the world. However, the general challenges of real-time reporting out of EHRs and separation between clinical teams and research, data and analytics are not particular to Sutter Health. We hope that others will share their own challenges and be inspired to work toward a more collaborative, transparent, and high-quality health data ecosystem.

Funding

None.

CRediT authorship contribution statement

Sylvia E.K. Sudat: Conceptualization, Writing - original draft, Writing - review & editing. **Sarah C. Robinson:** Writing - review & editing. **Satish Mudiganti:** Writing - review & editing. **Aravind Mani:** Writing - review & editing. **Alice R. Pressman:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Mageit, Scientists collaborate over new COVID-19 data portal, Healthcare IT News, 2020. https://www.healthcareitnews.com/news/scientists-collaborate-over-new-covid-19-data-portal. Published 8/27/2020 (accessed 10/20/2020).

[2] J. LoTempio, D.A. Spencer, R. Yarvitz, A. Delot Vilain, E. Vilain, E. Délot, We Can Do Better: Lessons Learned on Data Sharing in COVID-19 Pandemic Can Inform Future Outbreak Preparedness and Response, Sci. Diplomacy (2020) https://www.sciencediplomacy.org/article/2020/we-can-do-better-lessons-learned-data-sharing-in-covid-19-pandemic-can-inform-future. Published 8/19/2020 (accessed 10/20/2020).

[3] UKRI Ctse-, Why share scientific data during a pandemic? https://coronaviruse xplained.ukri.org/en/article/vdt0011/. Published 2020 (accessed 10/20/2020).

[4] N.G. Weiskopf, C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, J. Am. Med. Inform. Assoc. 20 (1) (2013) 144–151.

[5] N. Weiskopf, S. Bakken, G. Hripcsak, C. Weng, A Data Quality Assessment Guideline for Electronic Health Record Data Reuse, EGEMS (Washington, DC) 5 (1) (2017).

[6] M. Kahn, J. Brown, A. Chun, et al., Transparent reporting of data quality in distributed data networks, EGEMS (Washington DC) 3 (1) (2015).

[7] T. Callahan, A. Bauck, D. Bertoch, et al., A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks, EGEMS (Washington DC) 5 (1) (2017).

[8] K. Chan, J. Fowles, J. Weiner, Review: electronic health records and the reliability and validity of quality measures: a review of the literature, Med. Care Res. Rev.: MCRR 67 (5) (2010).

[9] M. Kahn, T. Callahan, J. Barnard, et al., A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data, EGEMS (Washington DC) 4 (1) (2016).

[10] M. Kahn, M. Raebel, J. Glanz, K. Riedlinger, J. Steiner, A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research, Med. Care 50 (Suppl(0)) (2012).

[11] S. Liaw, J. Guo, S. Ansari, et al., Quality assessment of real-world data repositories across the data life cycle: A literature review, J. Am. Med. Inform. Assoc.: JAMIA (2021).

[12] M. Reynolds, A. Bourke, N. Dreyer, Considerations when evaluating real-world data quality in the context of fitness for purpose, Pharmacoepidemiol. Drug Saf. 29 (10) (2020).

[13] J. Diaz-Garelli, E. Bernstam, M. Lee, K. Hwang, M. Rahbar, T. Johnson, DataGauge: A Practical Process for Systematically Designing and Implementing Quality Assessments of Repurposed Clinical Data, EGEMS (Washington DC) 7 (1) (2019).

[14] C. Lin, K. Bookman, A. Sieja, et al., Clinical Informatics Accelerates Health System Adaptation to the COVID-19 Pandemic: Examples from Colorado, J. Am. Med. Inform. Assoc.: JAMIA (2020).

[15] S. Nan, T. Tang, H. Feng, et al., A Computer-Interpretable Guideline for COVID-19: Rapid Development and Dissemination, JMIR Med. Inform. 8 (10) (2020).

[16] S. Deeds, S. Hagan, J. Geyer, et al., Leveraging an electronic health record note template to standardize screening and testing for COVID-19, Healthcare (Amsterdam Netherlands) 8 (3) (2020).

[17] J.J. Reeves, H.M. Hollandsworth, F.J. Torriani, et al., Rapid response to COVID-19: health informatics support for outbreak management in an academic health system, J. Am. Med. Inform. Assoc. 27 (6) (2020) 853–859.

[18] J. Wu, J. Wang, S. Nicholas, E. Maitland, Q. Fan, Application of Big Data Technology for COVID-19 Prevention and Control in China: Lessons and Recommendations, J. Med. Internet Res. 22 (10) (2020).

[19] C. Sáez, N. Romero, J. Conejero, J. García-Gómez, Potential limitations in COVID-19 machine learning due to data source variability: a case study in the nCov2019 dataset, J. Am. Med. Inform. Assoc.: JAMIA (2020).

[20] R. Biswas, A. Afiaz, S. Huq, Underreporting COVID-19: the curious case of the Indian subcontinent, Epidemiol. Infect. 148 (2020).

[21] J. Christianson, C. Carlin, L. Warrick, The dynamics of community health care consolidation: acquisition of physician practices, Milbank Quart. 92 (3) (2014).

[22] W. Weeks, R. Greene, J. Weinstein, Potential advantages of health system consolidation and integration, Am. J. Med. 128 (10) (2015).

[23] M. Furukawa, R. Machta, K. Barrett, et al., Landscape of Health Systems in the United States, Med. Care Res. Rev.: MCRR 77 (4) (2020).

[24] N. Beaulieu, L. Dafny, B. Landon, J. Dalton, I. Kuye, J. McWilliams, Changes in Quality of Care after Hospital Mergers and Acquisitions, New England J. Med. 382 (1) (2020).

[25] CHA COVID Tracking Tool: Data Dictionary. https://www.cdph.ca.gov/Progra ms/CHCQ/LCP/CDPH%20Document%20Library/AFL-20-31-Attachment-01.pdf. Published 2020 (accessed 1/3/2021).

[26] AFL 20-31. California Department of Public Health. https://www.cdph.ca.gov/Pr ograms/CHCQ/LCP/Pages/AFL-20-31.aspx?_cldee=bWJyeWFudEBjYWxob3NwaX RhbC5vcmc%3d&recipientid=contact-f311c9782f9fe811812b00155d0ca 31f-353f7b0bb4bf40b598f054e17bfe7bcb&esid=9d589683-edca-ea11-a 812-000d3a3abdcf. Published 2020 (accessed 10/20/2020).

[27] COVID-19 Hospital Data - California Open Data. https://data.ca.gov/dataset/covid -19-hospital-data. Published 2020 (accessed 10/20/2020).