HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2022 June 01.

Published in final edited form as:

J Biomed Inform. 2021 June ; 118: 103789. doi:10.1016/j.jbi.2021.103789.

Critical carE Database for Advanced Research (CEDAR): An Automated Method to Support Intensive Care Units with Electronic Health Record Data

Edward J. Schenck, MD, MS[1], **Katherine L. Hoffman, MS**[2], **Marika Cusick, MS**[3], **Joseph Kabariti, MS**[3], **Evan T. Sholle, MS**[3], **Thomas R. Campion Jr., PhD**[2,3,4,5]

[1]Weill Department of Medicine, Weill Cornell Medicine, New York, NY

[2]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY

[3]Information Technologies & Services Department, Weill Cornell Medicine, New York, NY;

[4]Department of Pediatrics, Weill Cornell Medicine, New York, NY

[5]Clinical & Translational Science Center, Weill Cornell Medicine, New York, NY

Abstract

Patients treated in an intensive care unit (ICU) are critically ill and require life-sustaining organ failure support. Existing critical care data resources are limited to a select number of institutions, contain only ICU data, and do not enable the study of local changes in care patterns. To address these limitations, we developed the Critical carE Database for Advanced Research (CEDAR), a method for automating extraction and transformation of data from an electronic health record (EHR) system. Compared to an existing gold standard of manually collected data at our institution, CEDAR was statistically similar in most measures, including patient demographics and sepsis-related organ failure assessment (SOFA) scores. Additionally, CEDAR automated data extraction obviated the need for manual collection of 550 variables. Critically, during the spring 2020 COVID-19 surge in New York City, a modified version of CEDAR supported pandemic response efforts, including clinical operations and research. Other academic medical centers may find value in using the CEDAR method to automate data extraction from EHR systems to support ICU activities.

Edward J. Schenck, MD, MS: Conceptualization, Resources, Funding Acquisition, Writing – Original Draft, Writing – Review & Editing
Katherine L. Hoffman, MS: Formal analysis, Conceptualization, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing
Marika Cusick, MS: Formal analysis, Visualization, Writing – Original Draft
Joseph Kabariti, MS: Software, Validation, Conceptualization, Writing – Original Draft, Writing – Review & Editing
Evan T. Sholle, MS: Conceptualization, Resources, Project administration, Writing – Original Draft
Thomas R. Campion, Jr., PhD: Conceptualization, Resources, Funding acquisition, Writing – Original Draft, Writing – Review & Editing

Declaration of interests
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.

Introduction

Patients treated in an intensive care unit (ICU) are critically ill and require life-sustaining organ failure support. They are at risk for progressive organ failure and death, and the optimal care of these patients may translate into significant health gains. ICUs are also among the most data-rich environments in the world of health care[1]. Integrating data in a critical care setting is very difficult, as patients generate over 10,000 discrete data points in any single day[2,3]. Rational collection and analysis of data at immense scale is both feasible and valuable[4]. The depth of phenotyping, number of participants, and standardization of data collection is critical to these efforts[4]. Efforts to understand and utilize data generated in the ICU are still evolving. Challenges include the integration of information from different systems and building a comprehensive platform to organize and structure many data phenotypes. Several analyses of large scale datasets have already unlocked novel patterns in data that have yielded insight into ICU diseases such as sepsis[5,6] and the acute respiratory distress syndrome (ARDS)[7].

The gold standard for ICU level data is Medical Information Mart in Critical Care MIMIC[11], which includes a rich database and reproducibility through the use of a code repository[12]. MIMIC is the basis for over 1200 studies across the world, with over 600 citations in 2019 alone. However, MIMIC data is limited to patients admitted to an ICU during the years 2001-2019 at Beth Israel Deaconess Medical Center (BIDMC), a quaternary care facility in Massachusetts affiliated with Harvard Medical School. MIMIC is also limited exclusively to the ICU, lacking data from care delivered before and after this care location, such as the emergency department (ED) and general medicine unit. Although efforts have been made to standardize and optimize data utilization through the code repository[12], MIMIC is not extensively curated and requires extensive cleaning and processing to analyze. In addition, because researchers who analyze MIMIC may likely not be familiar with practices, policies, and procedures at BIDMC, clinical heuristics concerning natural variation in care may be limited. Newer multicenter static databases such as the eICU Collaborative Research Database (eICU-CRD)[14] have been launched but are subject to similar limitations.

Secular trends in critical care, including the use of novel therapeutics, such as high-flow nasal canula, and treatment of novel diseases, such as COVID-19, continuously evolve, and practice patterns vary across institutions [13]. The COVID-19 pandemic caused a necessary unprecedented expansion of ICU care in New York City during the spring of 2020, which required emergent, near real time evaluation of practice patterns and clinical outcomes in order to inform decision making about resource utilization and staffing. A static retrospective database such as MIMIC would not have been able to adapt to these changing ICU definitions or have been useful during the crisis itself.

To support research and care, ICUs need a data collection method that is adaptable, accurate, and clinically useful in near real time[26,27]. Although standardized manual chart abstraction can enable an institution to understand local care practices within and beyond the ICU (e.g., ED, floor), limitations of the approach include arduous manual effort, risk of human error, potential subjectivity of chart abstraction, and the inability to scale across

thousands of patients. An alternative is automated extraction and transformation of data from the EHR to dynamically address ICU needs. To test the hypothesis that automated EHR data extraction could address limitations of existing ICU data collection approaches, we developed the Critical carE Database for Advanced Research (CEDAR). The purpose of this paper is to describe the design and evaluation of CEDAR as well as how CEDAR supported COVID-19 response efforts at our institution.

Methods

Setting

NewYork-Presbyterian/Weill Cornell Medical Center (NYP/WCMC) is an academic medical center in New York City located on the Upper East Side of Manhattan. NYP/WCMC has 862 beds and more than 1,600 attending physicians with faculty appointments at Weill Cornell Medical College of Cornell University. Each year NYP/WCMC treats more than two million patients, including over 310,000 emergency department cases and 5,000 ICU cases. ICU capacity includes 118 beds across dedicated cardiac, cardiothoracic surgery, medical, surgical, pediatric, and neurosurgical units. To document clinical care, NYP/WCMC clinicians began using the Allscripts® Sunrise Clinical Manager (SCM) EHR system in 2008.

Since 2014 investigators have prospectively consented patients admitted to any ICU at the institution to participate in a registry study involving collection of biospecimens and clinical data. For each participant, trained research coordinators manually abstracted 600 variables from the EHR system into REDCap case report forms. Designed by board certified pulmonary and critical care physicians, the REDCap project collected demographics, hospital timeline, laboratory and microbiology results, ventilatory parameters, APACHE-II evaluations, sequential organ failure assessment (SOFA) calculations, diagnoses and treatments throughout hospital stay, and final outcomes such as 28-day mortality and total ventilator-free days. Supplemental Appendix 1 contains the REDCap data dictionary codebook, which has expanded over time to include additional variables.

To complete chart abstraction of 600 variables for one patient, research coordinators typically required 3-4 hours. Attending physicians adjudicated each REDCap record created by research coordinators. Although coordinators received training and oversight from clinical leadership, attending physicians frequently observed errors and inconsistencies in REDCap records that required adjustment. Notably, computation of ventilator parameters, APACHE-II, and SOFA scores were particularly challenging because of their use of multiple, time-relative data points and formulas with stepwise computations. Further complicating data collection was frequent staff turnover, which required additional training of new research coordinators. Although faculty and staff improved data collection quality over time to establish an institutional registry for ICU analysis, the process remained time-consuming, and the challenge of scaling from hundreds to thousands of ICU patient cases persisted.

System design

As described below, we created CEDAR to support general ICU purposes and then modified it to address COVID-19 purposes. To extract, transform, and load EHR data for CEDAR, we used existing institutional informatics infrastructure based on Microsoft SQL Server 2016 [21]. Development occurred iteratively through collaboration of informatics staff, clinicians, and biostatisticians. Supplemental Appendix 2 contains detailed requirements collected from investigators and implemented by informatics personnel. The Institutional Review Board of Weill Cornell Medical College approved this study.

General ICU purposes—We included all data for patients with at least one ICU visit at the institution as defined by the patient having had a bed location in one of the ICUs as recorded in the EHR. As shown in Figure 1, CEDAR consisted of two sets of relational database tables called *basic* and *enhanced* that contained clinical details about ICU patients.

Based on the data models of MIMIC-III and eICU-CDR, CEDAR *basic* tables consisted of individual clinical data elements extracted from the EHR that enabled users to examine distributions of values (e.g., laboratory results) and potentially determine new measures. Stated differently, *basic* tables contained "raw" EHR data. In contrast to MIMIC-III and eICU-CDR, *basic* tables contained data from ICU and non-ICU encounters during a hospitalization rather than only ICU activity.

To address the lack of validated measures (e.g., SOFA scores) in MIMIC-III and eICU-CDR, *enhanced* tables consisted of pre-calculated values determined using data from *basic* tables. As described in Table 1, we performed five types of numerical calculations to transform *basic* values to create entries in *enhanced* tables. Examples of variables in *enhanced* tables included daily SOFA scores, daily intubation status and associated ventilatory parameters, worst laboratory results in time windows of interest defined as furthest from a threshold value [24], concern for infection criteria, and ventilator and organ failure free days across the hospital stay.

Enhanced tables contained variables repeated at different time points, such as APACHE-II scores [24] at the first, third, and seventh days of ICU stay determined using blood pressure, respiratory rates, temperature, and oxygenation among other *basic* measures to determine severity of a patient's condition and risk of mortality. *Enhanced* tables also contained SOFA scores [25] determined using similar types of calculations as well as rule-based approaches as illustrated in Figure 2.

COVID-19 purposes—At the onset of the pandemic, clinicians and biostatisticians observed that data in neither the EHR nor CEDAR adequately supported decision making. While the EHR provided up-to-the-minute patient-level data, it did not enable population-level aggregation of clinical observations. Similarly, although CEDAR contained raw and computed data for all critically ill patients, the data were not current. Notably, CEDAR contained data only for patients treated in ICUs as documented in the EHR, but surge conditions in New York City changed institutional treatment patterns for the critically ill.

On April 3, 2020 in response to clinician feedback, we modified CEDAR to define the start of an ICU encounter as the start of ventilation for a patient rather than the start of a patient having a bed location in an ICU. Additionally, if start of ventilation did not exist in the EHR, we used start of hospitalization to represent start of an ICU encounter. The rationale for the change was that the institution extended critical care treatment from dedicated ICUs to non-ICUs (e.g., operating rooms) to address an acute influx of patients presenting with COVID-19. During the COVID-19 outbreak, there was an unprecedented need for understanding the trajectories of COVID-19 patients, especially those who required, or would eventually require, a ventilator.

Evaluation

We evaluated CEDAR with respect to general ICU purposes and COVID-19 purposes. First, to evaluate general ICU purposes, we used manually collected data in REDCap case report forms (as described above) as a gold standard. We compared manually collected data from REDCap versus automated data from CEDAR using Wilcoxon rank-sum and chi-squared tests as appropriate. Specifically, we compared demographics and SOFA scores. For SOFA scores, we compared the raw scores for each individual SOFA sub-score—respiratory system, nervous system, cardiovascular system, liver, coagulation, kidneys—as well as the composite total SOFA score. For each score type, we calculated the Pearson's correlation coefficient between the two methods of calculation. Second, to evaluate COVID 19-purposes, we observed how modifications to CEDAR supported pandemic response efforts.

Results

Of the 600 variables collected per ICU case, CEDAR automated extraction of 550 variables (92%). Of the remaining 50 variables, manual [22] and semi-automated REDCap methods [23] enabled completion of data collection.

General ICU purposes

CEDAR required approximately 18 hours to generate, and we refreshed it once per month. For the 550 variables manually abstracted for 177 patients and 39,152 patients automated by CEDAR, characteristics were largely similar (Table 2). Outcomes differed between the two populations because manual chart abstraction focused on sicker patients in the medical ICU whereas automated EHR extraction with CEDAR included patients from all ICUs at the institution.

Using automated EHR data extraction, we determined that each ICU had different primary diagnoses (Table 3). Of note, septicemia was the most common primary diagnosis in the medical ICU.

As shown in Figure 2, SOFA score distributions for manual chart abstraction and automated EHR extraction did not differ statistically (p=.99). The median SOFA score for chart review-derived values was 7 [IQR 5-10], and the median SOFA score for automatically-derived SOFA values was 8 [IQR 5-10].

As shown in Table 4, the Pearson's correlation coefficient between the two methods of SOFA score calculation was 0.88 for the composite SOFA measure. Subscore correlation ranged from 0.61 to 0.96, with the lowest being the respiratory system (r=0.61) subscore and the highest being the coagulatory subscore (r=0.96).

Automated data in CEDAR differed from corresponding elements derived through manual chart review principally in that manual data collection occasionally involved divergence from a strict implementation of the 24-hour window for calculation of SOFA scores. In many instances, manual reviewers appear to have seen a clinically relevant value occur just outside of the window – for example, an elevated body temperature 24 hours and 15 minutes after admission to the ICU – and choose to include this value in the calculation of the pertinent SOFA subscore. Confronted with the same case, the automated technique adhered to a strict definition of which observations to include, and rejected the value as falling outside the defined temporal window.

COVID-19 purposes

After minor expansions to its inclusion criteria to address critical care treatment occurring in ICU and non-ICU settings at the institution, CEDAR rapidly met needs for both hospital operations and clinical research. CEDAR already contained calculations necessary to answer many clinical questions pertinent to COVID-19, and adjustment of timing of ICU encounters enabled CEDAR to reflect immediate changes in clinical care patterns. The COVID-19-specific version of CEDAR was identical in structure to CEDAR but adapted for the specific needs of caring for critically ill COVID-9 patients, such as daily aggregate summaries beginning at the time of hospital admission rather than ICU admission. Similarity of COVID-19-specific CEDAR enabled clinicians and biostatisticians familiar with CEDAR to immediately use the new resource to address pandemic needs. The modified CEDAR specific to COVID required 8 hours of run-time, and we refreshed it once per week.

Laboratory results and SOFA scores were invaluable for understanding the risk of intubation for hospital operations purposes. A decision tree was created for clinicians in the emergency department to evaluate the probability a patient would require intubation. This risk prediction model was immediately used to aid in decisions of which patients to transfer to the Javits Center, a New York City convention center designated by authorities as a temporary COVID-19 hospital, and presented at a WCMC Medical Grand Rounds during the height of the outbreak in April 2020. The laboratory results in CEDAR also highlighted that important COVID-19 laboratory tests, such as C-Reactive Protein and d-dimer, were not being ordered regularly by clinicians, allowing chiefs of practice to enact new guidelines, including creation of a COVID-19 laboratory bundle.

In addition to supporting patient care operations, COVID-19 CEDAR data enabled clinical research. During April-June 2020, faculty published several papers, including an important initial summary published in the New England Journal of Medicine [18]. Shortly thereafter, many of the laboratory results in COVID-19 CEDAR were used in a paper describing obesity as a potential risk factor for poorer outcomes in COVID-19 hospitalized patients in Annals of Internal Medicine[19], and a detailed description of ventilator parameters and

treatments of intubated COVID-19 patients was also published in the Annals of the American Thoracic Society[20].

Discussion

We developed and evaluated CEDAR, a method for ICUs to automate data collection from electronic health record systems that addresses local changes in care patterns and includes observations from ICU and other non-ICU hospital encounters. CEDAR includes a number of pre-calculated aggregate measures as well as the raw data required to compute them. Development of CEDAR required extensive collaboration between clinical subject matter experts and informatics specialists, including an iterative process of requirements-gathering, data engineering, and quality assurance. Data aggregated from the electronic health record into CEDAR showed concordance with data from manual chart review, indicating that the transformation constitutes a reasonably faithful representation of the true underlying patient state and can reduce manual data collection effort for staff. In contrast to static ICU data sets such as MIMIC, CEDAR enables assessment of rapidly changing secular trends, such as the COVID-19 pandemic. We have used CEDAR to support clinical and translational research at our institution, and our experience suggests that the method represents a novel approach toward reuse of electronic health record data to support critical care research and operations that other institutions may wish to replicate.

While existing resources for conducting critical care research have led to significant breakthroughs in the field, their temporally bound, de-identified, and ICU-specific nature has limited their utility in several, clinically meaningful ways. Users of MIMIC-III are limited in their ability to understand new therapeutic pathways and evaluate the impact of quality-based interventions. For example, changing O_2 or blood pressure targets may have a meaningful impact on overall survival and 30-day readmission, as could the use of high-flow nasal cannula in the adult population, new vasopressors (e.g. angiotensin II), or new protocols for oxygen targets in the ICU, weaning from mechanical ventilation, sedation, insulin therapy, or others. The methodology described here allows clinicians to regularly evaluate the impact of interventions such as these on quality of care, moving closer towards realizing the goals of the learning health care system. Specifically, CEDAR was integral to our institutional research and clinical response to the COVID-19 pandemic. By providing near-real-time research quality data, we were able to analyze our clinical experience quickly, resulting in impactful and practice changing publications. Our clinical staff utilized CEDAR derived clinical data summaries to discuss practice patterns and trends as the first surge of the pandemic evolved. We helped define risk factors for worsening respiratory failure within our specific population and disseminated those results to our faculty to inform practice.

In comparing the results of the automated calculations of commonly used metrics (e.g. APACHE, SOFA, and concern for infection), CEDAR replicated the results of manual chart review with satisfactory performance (Pearson’s correlation coefficient = 0.88), indicating that data transformed from the EHR are of sufficient quality to rely on the conclusions of research projects making use of CEDAR data in implementing new interventions at the point of care. Because of the concordance between manual and automated data collection, we have subsequently generated a de-identified version of CEDAR to support activities

preparatory to research and non-human subjects research. CEDAR also includes data gathered from beyond the ICU, including the emergency department and floor, enabling additional analyses that would not be feasible using a data set gathered only from critical care settings.

Multiple groups have acknowledged the limitations of the MIMIC-III data set and sought to develop competitors. Notably, the eICU-CRD data set seeks to address a primary limitation of MIMIC-III – its single-center status – by integrating data from multiple intensive care units in community hospitals. Similarly, researchers with the United Kingdom's National Health Services have described a methodology for linking local, proprietary EHR data with a national resource for research, the Intensive Care National Audit and Research Centre (ICNARC)'s Case Mix Programme (CMP) data set. While these efforts have contributed significantly to the research enterprise, the methodology described herein, whereby local, site-specific data is transformed to a specific, research-ready data set with collaboration between clinical experts, biostatisticians, and informatics personnel, represents a contrasting approach that enables analysis of secular trends, site-specific interventions, and other novel use cases otherwise unachievable through the use of de-identified, temporally-bound critical care data sets.

Several limitations exist with regard to CEDAR and its implementation, including required effort and replicability. The human effort required to instantiate this resource was significant, with subject matter experts from across differing disciplines meeting regularly to collaborate and address observed issues with data quality and completeness. A corollary limitation of this methodology is that the WC-CEDAR code base is specific to our institution's EHR system. Other institutions seeking to adapt this methodology will, of necessity, need to engage a similar group of diverse practitioners to replicate our results. However, the data model described herein and the overall methodology of working iteratively to define requirements, transform data, and conduct quality assurance testing are generalizable across institutions with the resources and staff to carry them out. To encourage generalizability across settings, future work will address instantiation of CEDAR using the Observational Medical Outcomes Partnership (OMOP) common data model, which academic medical centers have widely adopted. Finally, we refreshed CEDAR on a monthly and weekly schedule for general and COVID-19 purposes, respectively, due to internal resource constraints (i.e., needing to refresh other databases using institutional infrastructure). Although the refresh schedule met needs of clinicians and biostatisticians, future work will address more frequent updates, such as using streaming data and other technologies, to improve support for real-time clinical decision making.

The implementation of CEDAR at our institution has already enabled novel research projects that would otherwise be unfeasible. We contend that the methodology we describe here constitutes a novel and potentially impactful alternative to the reuse of temporally-bounded, de-identified, ICU-only data sets from differing institutions. By transforming local EHR data in this fashion, institutions can conduct research that would not be feasible with MIMIC-III or eICU-CRD data and better address their own site-specific research and quality improvement goals, especially during a pandemic such as COVID-19.

Conclusion

Seeking to develop an ICU-focused data resource sensitive to local practice changes and capable of regular refreshes, we built CEDAR through a collaboration between clinicians, biostatisticians, and informatics professionals including iterative requirement definition, data extraction, quality assurance and review. CEDAR features both ICU and non-ICU data, and includes a number of computed metrics of common concern for critical care research. Comparing these automatically calculated metrics to the results of manual chart review indicated that the data in CEDAR are sound and fit for reuse, and WC-CEDAR is currently contributing to research at our institution as well as COVID-19 clinical care. Other institutions may find value in replicating this methodology using data from their electronic health record systems.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This study received support from NewYork-Presbyterian Hospital (NYPH) and Weill Cornell Medical College (WCMC), including the Clinical and Translational Science Center (CTSC) (UL1TR002384) and Joint Clinical Trials Office (JCTO). The authors thank Steven Flores for his contributions.

Bibliography

1. Celi LA, Mark RG, Stone DJ, et al. Big data" in the intensive care unit. Closing the data loop. Am J Respir Crit Care Med 2013;187:1157–1160. doi:10.1164/rccm.201212-2311ED [PubMed: 23725609]
2. Iwashyna TJ, Liu V. What's so different about big data?. A primer for clinicians trained to think epidemiologically. Annals of the American Thoracic Society 2014;11:1130–1135. doi:10.1513/AnnalsATS.201405-185AS [PubMed: 25102315]
3. Obermeyer Z, Lee TH. Lost in Thought - The Limits of the Human Mind and the Future of Medicine. N Engl J Med 2017;377:1209–1211. doi:10.1056/NEJMp1705348 [PubMed: 28953443]
4. Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. Nat Med 2020;26:29–38. doi:10.1038/s41591-019-0727-5 [PubMed: 31932803]
5. Seymour CW, Kennedy JN, Wang S, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA 2019;321:2003–2017. doi:10.1001/jama.2019.5791 [PubMed: 31104070]
6. Knox DB, Lanspa MJ, Kuttler KG, et al. Phenotypic clusters within sepsis-associated multiple organ dysfunction syndrome. Intensive Care Med 2015;41:814–822. doi:10.1007/s00134-015-3764-7 [PubMed: 25851384]
7. Calfee CS, Delucchi K, Parsons PE, et al. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. Lancet Respir Med 2014;2:611–620. doi:10.1016/S2213-2600(14)70097-9 [PubMed: 24853585]
8. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med 2016;375:1216–1219. doi:10.1056/NEJMp1606181 [PubMed: 27682033]
9. Bhavani SV, Carey KA, Gilbert ER, et al. Identifying novel sepsis subphenotypes using temperature trajectories. Am J Respir Crit Care Med 2019;200:327–335. doi:10.1164/rccm.201806-1197OC [PubMed: 30789749]
10. Hotchkiss RS, Sherwood ER. Getting sepsis therapy right. Science 2015.
11. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035. doi:10.1038/sdata.2016.35 [PubMed: 27219127]

12. Johnson AE, Stone DJ, Celi LA, et al. The MIMIC Code Repository: enabling reproducibility in critical care research. J Am Med Inform Assoc 2018;25:32–39. doi:10.1093/jamia/ocx084 [PubMed: 29036464]
13. Admon AJ, Seymour CW, Gershengorn HB, et al. Hospital-level variation in ICU admission and critical care procedures for patients hospitalized for pulmonary embolism. Chest 2014;146:1452–1461. doi:10.1378/chest.14-0059 [PubMed: 24992579]
14. Pollard TJ, Johnson AEW, Raffa JD, et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018;5:180178. doi:10.1038/sdata.2018.178 [PubMed: 30204154]
15. Ibrahim ZM, Wu H, Hamoud A, et al. On classifying sepsis heterogeneity in the ICU: insight using machine learning. J Am Med Inform Assoc Published Online First: 17 1 2020. doi:10.1093/jamia/ocz211
16. Pirracchio R, Petersen ML, Carone M, et al. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med 2015;3:42–52. doi:10.1016/S2213-2600(14)70239-5 [PubMed: 25466337]
17. Liu R, Greenstein JL, Granite SJ, et al. Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU. Sci Rep 2019;9:6145. doi:10.1038/s41598-019-42637-5 [PubMed: 30992534]
18. Goyal P, Choi JJ, Pinheiro LC, et al. Clinical characteristics of Covid-19 in New York City. N Engl J Med. 2020 6 11;382(24):2372–2374. doi: 10.1056/NEJMc2010419. Epub 2020 Apr 17. [PubMed: 32302078]
19. Goyal P, Ringel JB, Rajan M, et al. Obesity and COVID-19 in New York City: a retrospective cohort study. Ann Intern Med. 2020 7 6:M20–2730. doi: 10.7326/M20-2730.
20. Schenck EJ, Hoffman K, Goyal P, et al. Respiratory mechanics and gas exchange in COVID-19-associated respiratory failure. Ann Am Thorac Soc. 2020 9;17(9):1158–1161. doi: 10.1513/AnnalsATS.202005-427RL. [PubMed: 32432896]
21. Sholle ET, Kabariti J, Johnson SB, et al. Secondary use of patients' electronic records (SUPER): an approach for meeting specific data needs of clinical and translational researchers. AMIA Annu Symp Proc. 2018 4 16;2017:1581–1588. eCollection 2017. [PubMed: 29854228]
22. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 4;42(2):377–81. doi: 10.1016/j.jbi.2008.08.010. Epub 2008 Sep 30. [PubMed: 18929686]
23. Campion TR Jr, Sholle ET, Davila MA Jr. Generalize middleware to support of REDCap dynamic data pull for integrating clinical and research data. AMIA Jt Summits Transl Sci Proc. 2017 7 26;2017:76–81. eCollection 2017. [PubMed: 28815111]
24. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985 10;13(10):818–29. [PubMed: 3928249]
25. Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med. 1996 7;22(7):707–10. [PubMed: 8844239]
26. Mann JK, Kaffashi F, Vandendriessche B, Jacono FJ, Loparo K., 2020. Data Collection and Analysis in the ICU. In Neurocritical Care Informatics (pp. 111–134). Springer, Berlin, Heidelberg.
27. Sun Y, Guo F, Kaffashi F, Jacono FJ, DeGeorgia M. Loparo KA. INSMA: An integrated system for multimodal data acquisition and analysis in the intensive care unit. J Biomed Inform. 2020 6;106:103434. doi: 10.1016/j.jbi.2020.103434. Epub 2020 Apr 28. [PubMed: 32360265]

- To test the hypothesis that automated electronic health record (HER) data extraction could address limitations of existing intensive care unit (ICU) data collection approaches, we developed the Critical carE Database for Advanced Research (CEDAR).
- Compared to an existing gold standard of manually collected data at our institution, CEDAR was statistically similar in most measures, and it reduced data collection time for more than 550 variables.
- During the spring 2020 COVID-19 surge in New York City, a modified version of CEDAR supported pandemic response efforts, including clinical operations and research.

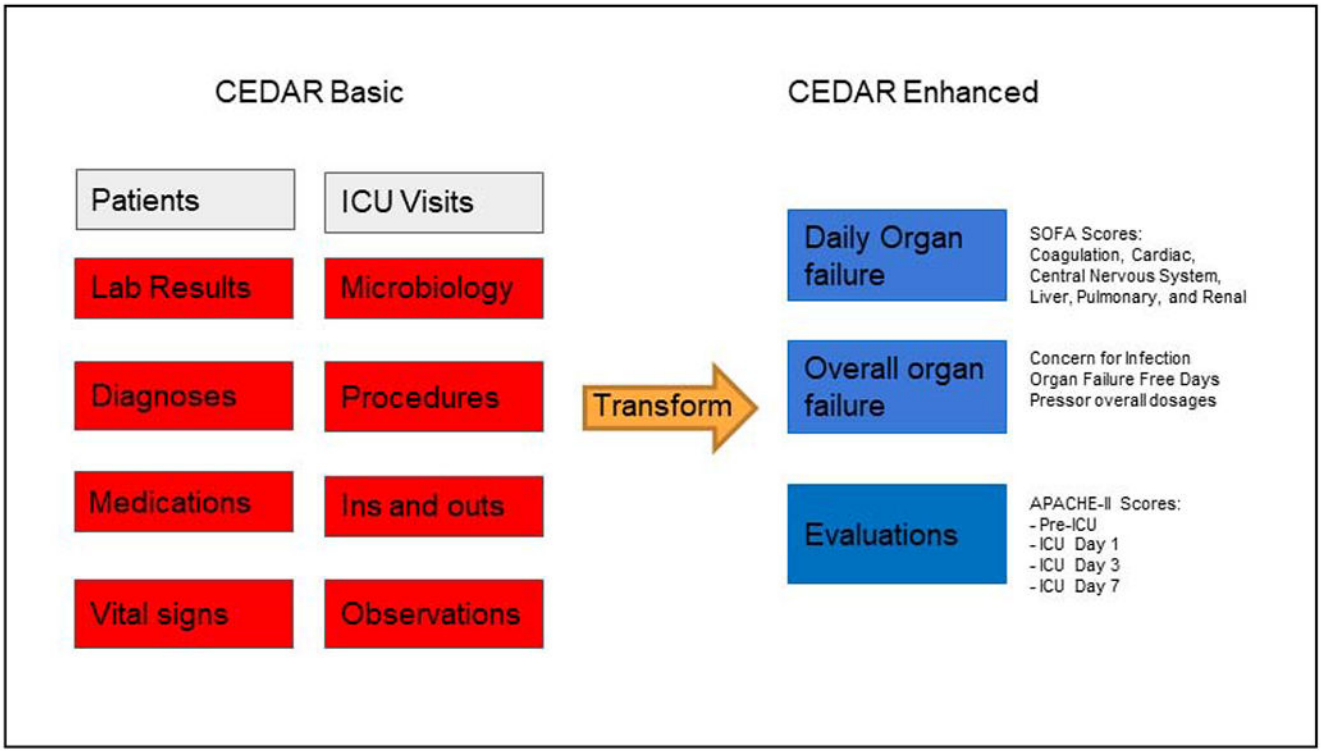

Figure 1.
CEDAR basic and enhanced relational database tables.

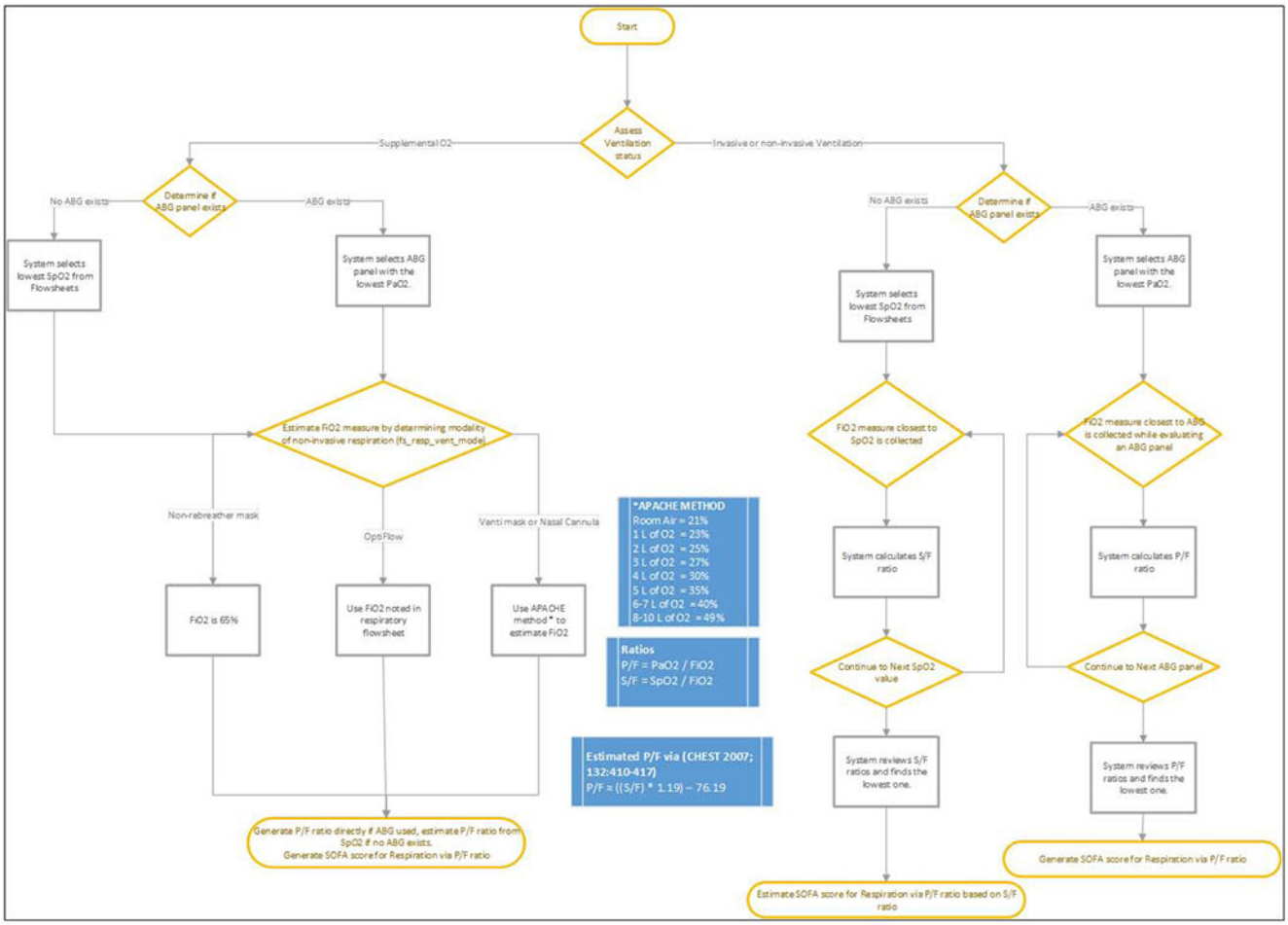

Figure 2.
Rule-based approach for SOFA score generation in *enhanced* table.

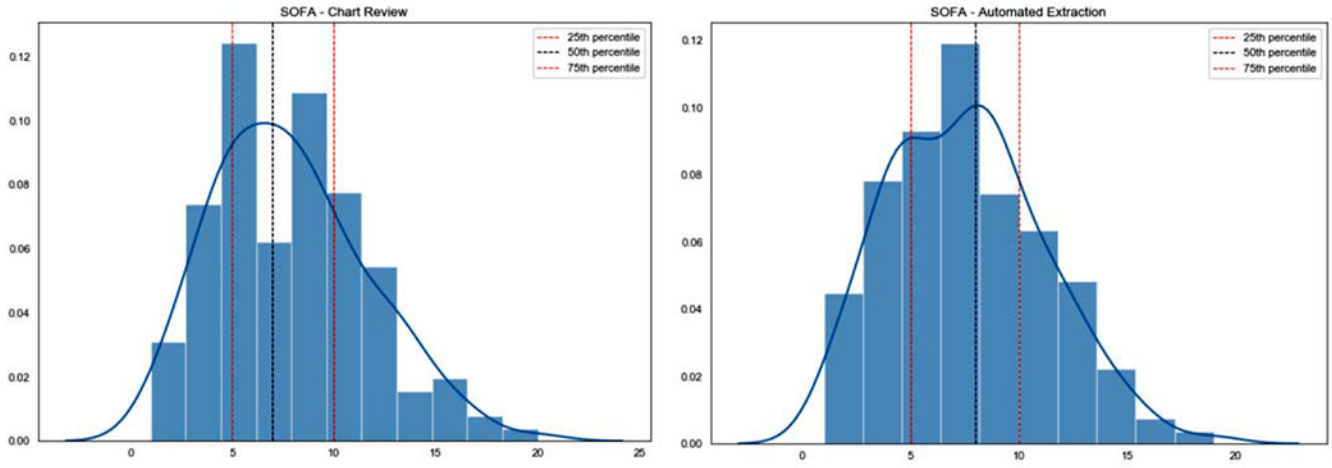

Figure 3.
SOFA distributions (Chart review, automated extraction)

Table 1.

Types of calculations performed for CEDAR *enhanced* tables.

Calculation type	**Calculated variable example**
Highest/lowest	Temperature, Heart Rate, Respiratory Rate, eGFR
Sum	IV fluids, Urine Output, Blood Transfusions
Closest to a particular time of day (e.g., 08:00)	Height, Weight, BMI, Urine Protein, Creatinine, Urine Osmolality
Furthest from threshold value	Hemoglobin (e.g., highest or lowest relative to 14 g/dL [24]), Sodium, Potassium, systolic blood pressure
Worst ABG and related components	PaO_2, $PaCO_2$, Ventilation status, Tidal Volume

Table 2.

Demographic characteristics of manual chart review and automated EHR extraction.

Characteristic	Manual Chart Abstraction (N=177)	Automated EHR Extraction (N=39,152)	p-value
Age	64 [51-75]	68 [52-80]	0.019
Female	83 (47%)	17,077 (44%)	0.4
Race			0.3
Asian/Indian	6 (3.4%)	2,365 (6.0%)	
Black	17 (9.6%)	3,659 (9.3%)	
Declined/Unknown/Other	78 (44%)	18,386 (47%)	
White	76 (43%)	14,742 (38%)	
BMI	27 [23-31]	26 [23-31]	0.3
ICU Length of Stay (Days)	5 [3-9]	4 [2-7]	<0.001
28-day Mortality	39 (23%)	3,240 (8.3%)	<0.001

Statistics presented: median [interquartile range] and N (%).

Table 3.

Primary diagnosis by ICU.

ICU	Primary Diagnosis	N (%)	Total
Cardiac ICU	Subendocardial infarction	1,184 (15.7%)	7,546
	Coronary atherosclerosis of native coronary artery	382 (5.1%)	
Cardiothoracic Surgery ICU	Aortic valve disorders	3,130 (38.4%)	8,145
	Coronary atherosclerosis of native coronary artery	1,145 (14.1%)	
Medical ICU	Septicemia	857 (10.9%)	7,829
	Acute respiratory failure	338 (4.3%)	
Mixed Surgical ICU	Cerebral aneurysm, non-ruptured	522 (3.24%)	16,115
	Benign neoplasm of cerebral meninges	181 (1.12%)	
Pediatric ICU	Compression of brain	75 (4.6%)	1,638
	Septicemia	289 (1.79%)	

Table 4.

Correlation by SOFA score type.

SOFA Score Type	Pearson's correlation coefficient
Coagulation	0.96
Liver	0.95
Composite	0.88
Cardiovascular system	0.87
Kidneys	0.81
Nervous system	0.78
Respiratory system	0.61