Validation of an Intensive Care Unit Data Mart for Research and Quality Improvement

Christina Boncyk (christina.s.boncyk@vumc.org)
Vanderbilt University Medical Center
Pamela Butler
Vanderbilt University Medical Center
Karen McCarthy
Vanderbilt University Medical Center
Robert E. Freundlich
Vanderbilt University Medical Center

Short Report

Keywords: electronic health record, intensive care unit, quality improvement, ICU, data mart

Posted Date: June 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1666890/v1

Abstract

Data derived from the electronic health record (EHR) frequently undergo limited prospective validation. Efforts to perform validation often suffer from limited collaboration between clinicians and data analysts. In this manuscript, we describe the creation of a structured, rigorously validated intensive care unit (ICU) data mart based on data automatically and routinely derived from the EHR. This data mart includes high-quality outcomes data and other data elements commonly used for quality improvement and research purposes. These data elements were identified by physicians working closely with data analysts to iteratively develop and refine algorithmic definitions for complex outcomes and risk factors. We contend that this methodology can be reproduced and applied to other clinical domains to create high quality data marts, inclusive of complex outcomes data.

Background

The electronic health record (EHR) contains a vast quantity of data due to its observation nature, holding great promise as a valuable, efficient, and cost-effective tool. These data can inform quality improvement and research initiatives, especially those related to medical resources and patient outcomes.[1–3]. In its initial implementation, however, the EHR rarely captures outcomes of interest to key stakeholders reliably and accurately due to frequent limitations resulting from disorganized, incorrect, or missing variables that lack vigorous extraction methodologies. Together this limits the data's validity and utility.[4] In order to provide validated results for scientific interpretation, vigorous, reproducible, and validated techniques must be established for each EHR variable of interest.

Many institutions rely on a structured repository of data, drawn from the EHR, to facilitate ongoing access, a so-called data warehouse or data mart.[5, 6] This data repository is frequently created after the EHR has been created and, at many institutions, is created and maintained by data analysts working in isolation from front-line clinicians. The intensive care unit (ICU) is a particularly challenging area for creation of a data mart. Critically ill patients suffer life threatening organ pathology in at least one, if not many, organ systems. These patients are generally intensively monitored, with very high frequency of physiologic data capture. Laboratory data may be obtained multiple times per day. Multiple organ support modalities may be employed, with complex documentation and monitoring to quantify the degree of support. Once data are located though, they can support surveillance, decision support, and modeling of outcomes.[7]

We sought to define a methodology for the creation of a structured, rigorously validated intensive care unit (ICU) data mart based on data automatically and routinely derived from the EHR. We identified data elements commonly used for quality improvement and research purposes, including high-quality outcomes data, to refine the methodologies presented here. Importantly, data analysts and clinicians worked side-by-side throughout the process.

Methods

A multidisciplinary project team including clinical intensivists and data analysts met daily throughout construction, assembling a priority list of physiologic, laboratory, demographic, and billing data. The presence of data availability in routine clinical practice was confirmed using extensive chart review and validation. Data analysts worked to identify the location of variables within the data architecture underlying the EHR (Epic, Verona, WI). All elements were investigated for extent of chart documentation, frequency, duplication in different sites within the medical record, and agreement with patient clinical course. Variable extraction was initiated after location(s) of these key variables within the EHR were confirmed.

We created algorithmic definitions for complex data elements, including most outcomes, leveraging existing literature, when available. Test patients were extracted and algorithms iteratively refined at least weekly. The number of iterations required for each variable was variable dependent, influenced by fidelity of the variable within the EHR. Once shown to be reproducible in a broad cohort of patients, structured query language (SQL) was used to extract, transform, and load data from the EHR into a relational database housed on a departmental server. This same departmental server houses intraoperative variables within a perioperative data warehouse (PDW) that were obtained through similar methodologies.[8] The sensitivity and specificity of algorithmic definitions were then formally assessed using the finalized algorithms on random cohorts of 50–100 ICU patients. After formal assessments, patient outcome results were cataloged in tables for presentation and dissemination. This process is illustrated in Fig. 1.

Figure 1 Variables presented have been identified by clinicians as high-yield variables for quality improvement and research purposes. As these variables are located and validated within the electronic health record, they are added within the data mart. These data are then analyzable to be able to draw conclusive findings regarding outcomes including acute kidney injury, reintubation rates, and 7-day mortality, to name a few

Results

A total of 459,465 ICU patient encounters were identified and included in the ICU data mart. These patients include over 460,000,000 individual laboratory results and 4,610,776 vital signs (with 1-minute fidelity in the first 24-hours of admission). Using the above methodologies, a total of 26 outcomes were validated (Table 1). These data have been structured within 19 tables, all of which have a sensitivity and specificity of greater than 95%. The iterative construction of our processes allows for continual updates and validation of variables to maintain accuracy. When variables were not able to reach sufficient sensitivity or specificity despite iterations, data analysts formed collaborative meetings to review, check, and improve techniques. If it wasn’t possible to increase accuracy, variables were not advanced or included within final tables or projects due to lack of validity.

Table 1
Variables Currently Validated Organized by Variable Type

Variable Type	Variables
Patient Variables	Date of birth, Height, Weight, BMI
	Ethnicity, Gender, Sex, Race
	Date of surgery/anesthesia
	Primary admission diagnosis
	Insurance type, Smoker History
	Mortality scores: Charlson, Elixhauser, Romano
	ICU discharge location
	Hospital admission and discharge dates
	ICU admission and discharge dates
	Readmission date within 7 days
	ED visits within 7 days of discharge
	Death date
Laboratory/Imaging Variables	Approximately 4100 unique lab test variables (e.g., BUN, drug levels)
	Hematocrit values
	INR values
	Partial thromboplastin values
	WBC count min/max
	CT head scans
	Pathology results

Variable Type	Variables
Medication Variables	Crystalloid IV fluids (normal saline, LR, plasmalyte)
	575 unique transfusions
	Albumin
	Parenteral nutrition
	Enteral nutrition
	Blood transfusion (pRBCs, FFP, platelets, cryoprecipitate)
	750 Infused medications, including nutrition (e.g., electrolytes, dextrose,), diabetic control (e.g., insulin, dialysates, etc) and medications (e.g., fentanyl, propofol, dexmed)
	400 Antibiotics (e.g., ampicillin, vancomycin, cefazolin)

Variable Type	Variables
Hospital Course Variables	First recorded time of dialysis treatment:
	CRRT, HD, PD
	Mechanical Ventilator variables:
	Ventilator Mode, SpO2, FiO2
	Mandatory Respiratory Rate, SpO2r
	Tidal (Observed), PEEP/CPAP (cm H2O)
	Central venous pressure values
	Assessment scores:
	GCS, SOFA, RASS, CAM-ICU, Peds NLP score
	Peds PEW score
	Arterial pressure min/max
	Pulse min/max
	Systolic BP min/max
	Temp min/max
	Urine Output min/max
	Central Line and Cath durations
	Chest effort flag
	Wheeze flag
	ECMO RPM
	O2 lpm
	Intubation duration
	Extubation time
	Reintubation date

Validated data are used to interpret patient outcomes for all patients within the ICU and ICU data mart. Additionally, these data can be joined to the 120 tables including more than 1900 unique variable columns within the existing perioperative data warehouse.

Discussion

We present a methodology for building a robust and highly granular ICU data mart, leveraging the synergistic expertise of clinicians and data analysts. Optimizing the quality of data obtained from large databases will improve accuracy, results, and confidence within informatics research for quality improvement and research purposes.

These processes can be adapted to new variables as they present to provide real-time clinical data on large populations of patients within our ever-changing clinical environment. Several hospital systems involved in data informatics research have already established similar organizational methodologies to ensure quality of data obtained within their data warehouses.[6, 9, 10] Together, these structured and validated methodologies strengthen the results obtained and the validity and trust within our research community. Even after establishing these methodologies, there is a need for consistent upkeep and maintenance of systems. Continual data maintenance and validation is not included within these methods but are equally essential to ensuring continuation of valid data collection. The major value for this established methodology lies in the additional variables and patient markers that are added to the EHR and identified as priority for inclusion within the data mart. These same processes are adapted to ensure quality data collection and trust of information obtained. Within our workgroups, we are validating the accuracy of a variable or variables that will identify positive coronavirus 2019 (COVID-19) test results into our variable lists using the presented methodology to confirm accuracy within results obtained across a variety of available laboratory data.

Similar to much of informatics research, our results are limited by the quality of data entered into the EHR. Missingness and inaccurate data elements can be screened and eliminated when detected, but such errors are difficult to prevent entirely. The ability of our algorithms and methods to detect data accuracy with sensitivity and specificity > 95% on repeated queries is evidence of the rigor of our data extraction method. We recognize, however, this number is not 100% and thus a low level of inaccuracies cannot be eliminated. As our systems change and update within our underlying EHR architecture, aspects of our data extraction and validation may need to be updated as well to ensure continued validity.

Our methodology and accuracy provides a strong foundation for the results obtained through our large ICU data mart. As we plan to add patient data throughout the hospitalization and perioperative periods, we will continue to establish structured methodologies to ensure data accuracy. Future uses of these formats will aim to target validation of variables across multiple health care centers to create multicenter perioperative data warehouses with validated patient variables for quality improvement and research purposes.

Declarations

Conflict of Interest: The authors declare no competing interests.

Clinical Trial Number: Not applicable

Acknowledgments: Not applicable

Compliance with Ethical Standards

Conflict of interest statement: Dr. Boncyk is a consultant for Sedana Medical. This manuscript does not reference any activities related to this consultancy and Dr. Boncyk declares no conflict of interest. Dr. Freundlich declares no conflict of interest. Pamela Butler declares no conflict of interest. Karen McCarthy declares no conflict of interest.

Role of funding source: Dr. Boncyk receives support from the National Institute on Aging (NIA) Administrative Supplement (R01AG053582), the Foundation for Anesthesia Education and Research (FAER), and the Society of Academic Associations of Anesthesiology & Perioperative Medicine (SAAAPM). Dr. Freundlich is supported by two grants from the National Institutes of Health (NIH): National Heart, Lung, and Blood Institute (NHLBI) grant K23HL148640 and National Center for Advancing Translational Sciences (NCATS) grant UL1TR002243. There was no direct funding towards this project.

Ethical Approval: This project received local Intstitutional Review Board (IRB) approval (#220897).

Informed Consent: A waiver of informed consent was obtained with IRB approval.

Authors Contributions: CB contributed to project design, data interpretation, and construction of manuscript. PB contributed to data collection and construction of manuscript. KM contributed to project design, data collection, data interpretation, and construction of manuscript. RF contributed to project design, data collection, data interpretation, and construction of manuscript.

References

1. Brundin-Mather R, Soo A, Zuege DJ, Niven DJ, Fiest K, Doig CJ, et al. (2018) Secondary EMR data for quality improvement and research: A comparison of manual and electronic data collection from an integrated critical care electronic medical record system. J Crit Care. 47:295–301.
2. Chen LM, Kennedy EH, Sales A, Hofer TP. (2013) Use of health IT for higher-value critical care. N Engl J Med. 368:594–597.
3. King J, Patel V, Jamoom EW, Furukawa MF. (2014) Clinical benefits of electronic health record use: national findings. Health Serv Res. 49:392–404.
4. Docherty AB, Lone NI. (2015) Exploiting big data for critical care research. Curr Opin Crit Care. 21:467–472.
5. Kimball R, Ross M, Thorthwaite W, Becker B, Mundy J. The data warehouse lifecycle toolkit: John Wiley & Sons; 2008.
6. de Mul M, Alons P, van der Velde P, Konings I, Bakker J, Hazelzet J. (2012) Development of a clinical data warehouse from an intensive care clinical information system. Comput Methods Programs Biomed. 105:22–30.
7. Herasevich V, Pickering BW, Dong Y, Peters SG, Gajic O. (2010) Informatics infrastructure for syndrome surveillance, decision support, reporting, and modeling of critical illness. Mayo Clin Proc.

85:247–254.

8. Hofer IS, Gabel E, Pfeffer M, Mahbouba M, Mahajan A. (2016) A systematic approach to creation of a perioperative data warehouse. Anesth Analg. 122:1880–4.
9. Dewitt JG, Hampton PM. (2005) Development of a data warehouse at an academic health system: knowing a place for the first time. Acad Med. 80:1019–1025.
10. Weir CR, Hicken BL, Rappaport HS, Nebeker JR. (2006) Crossing the quality chasm: the role of information technology departments. Am J Med Qual. 21:382–393.

Figures

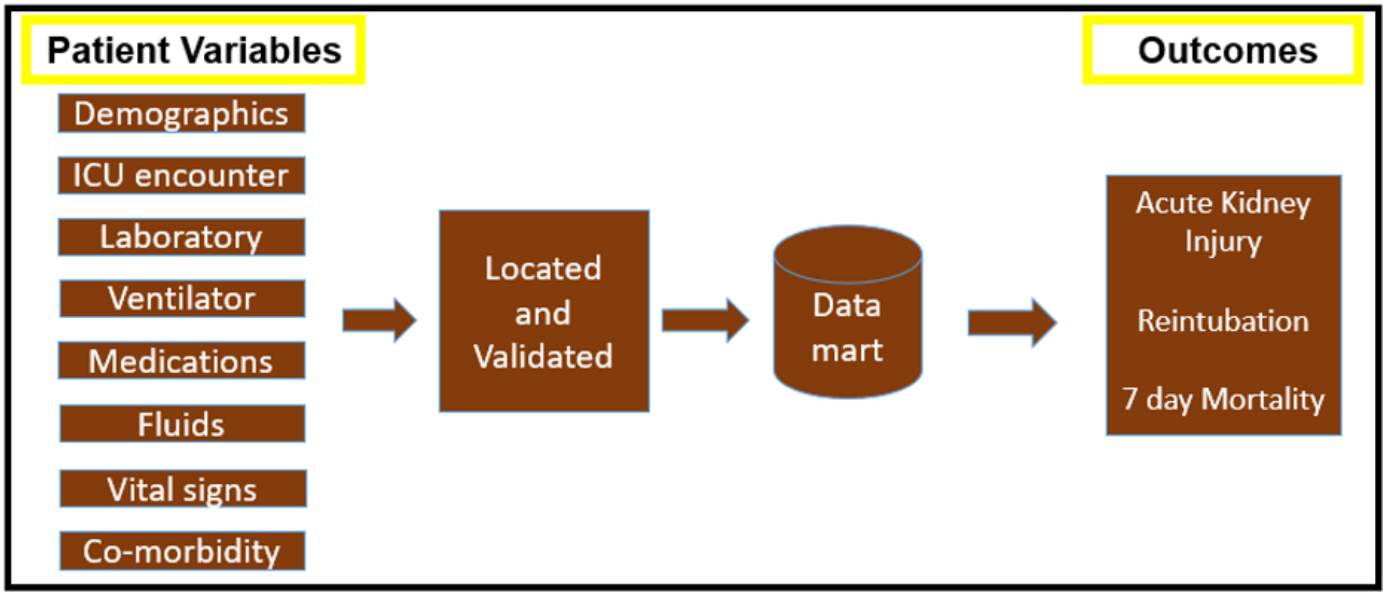

Figure 1

Progression Through Variable Identification or Validation of Patient Outcomes