Validating an Electronic Health Record (EHR) Phenotype of Patients Hospitalized for COVID-19

Lauren Heery, BS[1]; Kristine Erlandson, MD[1]; Krithika Suresh, PhD[2,3]; Will Carter, MIS[3]; Lisa Schilling, MD[4]

1. Division of Infectious Diseases 2. CSPH Department of Biostatistics and Informatics 3. ACCORDS 4. Division of General Internal Medicine

University of Colorado Anschutz Medical Campus

Colorado Clinical and Translational Sciences Institute (CCTSI)
UNIVERSITY OF COLORADO DENVER | ANSCHUTZ MEDICAL CAMPUS

Background

- As of April 9, 2021, there have been 31,023,288 confirmed cases of COVID-19 and 560,315 deaths in the U.S. (1)
- Observational EHR-derived data has been increasingly used to characterize patient-level aspects of the pandemic including the natural history of COVID-19, the impact of therapeutic interventions, and outcomes of healthcare systems (2) providing real-time clinical knowledge
- Despite the widespread use of these methods, it is often unknown the extent to which this data is valid and high quality (3)

Rationale: This study compares an EHR-computed phenotype of patients hospitalized for COVID-19 with a database of manually abstracted patient charts.

Methods

- **REDCap chart review database** of patients hospitalized for COVID-19 March 18-April 26, 2020 at UCHealth University of Colorado Hospital; inclusion based on COVID-19 infection flag in Epic, first admission
- The **EHR-computed phenotype** was developed through the Rapid Response Data for Discoveries (R2D2) collaboration and based on the Common Data Model (4)
- A query was designed to match the REDCap database cohort and search data in Health Data Compass (HDC) (5)
- Date of birth, age, race, ethnicity, gender self-reported by patients in the EHR
- Clinical variables of interest were abstracted and extracted from Epic charts as reported in flowsheets and notes
- Patients in the HDC dataset were linked by an arbitrary identifier with the patients in the REDCap database
- These patient records were manually compared to determine validity of the phenotype and query of HDC

Inclusion Criteria for REDCap Database and HDC Dataset

Criteria	REDCap Database	HDC Dataset
Patient population	Hospitalized adults	Hospitalized adults (needed visit type of inpatient admission)
Clinical sites	U of Colorado Hospital	UCH all University and AMC sites
Admission dates	3/18/20 - 4/26/20	3/18/20 - 4/26/20
COVID-19 case definition	Infection flag in Epic (Infection Control Team generated a patient list)	Positive COVID-19 test associated with encounter OR COVID-19 diagnosis code

Results

Figure 1: EHR-computed phenotype and chart review database comparison

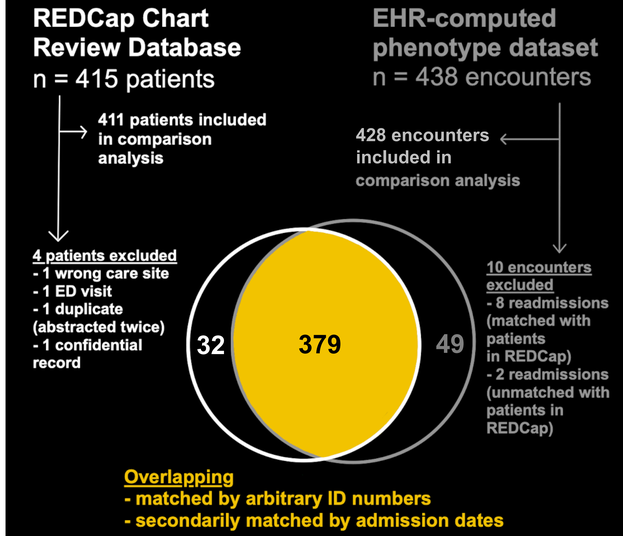

Table 1: Patient encounters identified by chart review and missed by EHR-computed phenotype (n = 32)

Incorrect visit type in data warehouse (inpatient admission coded as ED visit)	Visit information error in data warehouse (COVID-19 diagnosis code not linked with admission)	Phenotype criteria excluded (true cases excluded by date-restricted diagnosis codes)	Incorrect visit type in data warehouse and phenotype criteria exclusion	Not included in analysis
18	9	2	1	2

Table 2: Percent Accordance of Demographics Information in Overlapping Records

	Race	Ethnicity	Gender	Age
n = 379	92.3%	98.9%	99.7%	99.2%

Table 3: Comparison of Encounter Procedures and Outcomes in Overlapping Records

Records in HDC	Mechanical Ventilation (n=148)	Extracorporeal Membrane Oxygenation (n=10)	Deceased at discharge (n=121)
Overlapping Records (n)	115	10	119
% matched occurrence	77.7%	100%	98.3%
% matched dates	63.5%	n/a	n/a

Conclusions

- An EHR-computed phenotype had a 92.2% sensitivity compared with chart review in identifying cases of patients hospitalized for COVID-19.
- Analysis of patients missed by the EHR phenotype identified errors in data extraction, transformation and loading (ETL), particularly with regard to visit type in patients' encounters.
- This work highlighted limitations of applying date-restricted logic in the phenotype definition to ICD-10 diagnosis codes. It also showed the limitations of defining intensive services as only requiring mechanical ventilation or ECMO.
- Demographic information reliably underwent ETL from the EHR into HDC.
- Comparison of patient records highlighted problems with how procedures such as intubation undergo ETL in HDC.

Implications and Future Work

- This work has led to improvements in the phenotype definition's inclusion criteria and logic, and HDC's process.
- Data quality validation is critical as EHR-derived data and phenotypes are increasingly used in observational research.
- These findings ensure integrity of this EHR-data and strengthens the reliability of the Common Data Model.

References 1. WHO COVID-19 Dashboard. Geneva: World Health Organization, 2020. Available online: https://covid19.who.int/ (last cited: April 9, 2021). 2. Haendel MA, Chute CG, Bennett TD, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. 2021;28(3):427-443. doi:10.1093/jamia/ocaa196 3.Callahan TJ, Bauck AE, Bertoch D, et al. A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks. EGEMS (Wash DC). 2017;5(1):8. Published 2017 Jun 12. doi:10.5334/egems.223 4. Boyce RD, Ryan PB, Norén GN, et al. Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest. Drug Saf. 2014;37(8):557-567. doi:10.1007/s40264-014-0189-0 5. Rapid Response Data for Discoveries (R2D2) Collaboration https://covid19questions.org/

Acknowledgements Thanks to REDCap chart review team: Lyndsey Babcock, Connor Fling, Kellen Hirsch, Dave Sheneman, Nemanja Vukovic, and Taylor Wand. Thanks to Health Data Compass team: Ufi Olakpe, Michelle Edelmann, Michael Kahn, and Ian Brooks. Supported by NIH/NCATS Colorado CTSA Grant Number TL1 TR002533 and NHLBI GEMS Program Grant Number 5R25HL103286-10. Contents are the authors' sole responsibility and do not necessarily represent official NIH views. Supported by Health Data Compass Data Warehouse project (healthdatacompass.org). REDCap supported by NIH/NCATS Colorado CTSA Grant Number UL1 TR002535. Contents are the authors' sole responsibility and do not necessarily represent official NIH views. The Rapid Response Data for Discoveries (R2D2) Collaboration is funded by the Gordon and Betty Moore Foundation.

Disclosures WC, LH, LS, KS: nothing to disclose. KE: I have received research funding (to the University of Colorado) from Gilead Sciences, and consultant payment from Viiv Pharmaceuticals and Theratechnologies (to the University of Colorado).