Journal of the American Medical Informatics Association, 27(9), 2020, 1341–1342
doi: 10.1093/jamia/ocaa187
Editorial

Editorial

Informatics impact requires effective, scalable tools and standards-based infrastructure

Suzanne Bakken

School of Nursing, Department of Biomedical Informatics, and Data Science Institute, Columbia University, New York, New York, USA

Corresponding Author: Suzanne Bakken, PhD, RN, School of Nursing, Department of Biomedical Informatics, and Data Science Institute, Columbia University, 630 W. 168th Street, New York, NY 10032, USA; sbh22@cumc.columbia.edu

Received 23 July 2020; Accepted 23 July 2020

Our field of biomedical and health informatics frequently considers the question of the impact of our research and its application on discovery, care delivery, health, and health equity. In this editorial I highlight 5 papers that illustrate scalable tools and standards-based infrastructure. Such innovations are components of the essential foundation for realizing the impact of informatics. Two papers address the critical issue of privacy preservation in secondary use of electronic health record (EHR) data,[1,2] while a third focuses on information retrieval for COVID-19-related questions.[3] Two additional papers focus on data definitions, codified vocabularies, and other components that enable semantic interoperability and their particular role in the COVID-19 pandemic.[4,5]

Carrell et al describe the hiding in plain sight (HIPS) approach which replaces personally identifying information (PII) tagged by a deidentification system with resynthesized content with the intent of making it harder to detect unredacted PPI.[1] They used 2000 representative clinical documents from each of 2 healthcare settings to generate 2 deidentified 100-document corpora (200 documents total) where PII tagged by a typical automated machine-learned tagger was replaced by HIPS resynthesized content with a 10% leaky PII rate. Two readers from the originating institution and 2 external readers conducted aggressive reidentification attacks to isolate leaked PII. Mean recall and precision, respectively, varied for patient ages (9%, 37%), dates (32%, 26%), doctor names (25%, 37%), organization names (45%, 55%), and patient names (23%, 57%). Both recall and precision were higher for internal than external readers. While the HIPS results were superior to published findings of traditional redaction, they were inferior to a human adversary augmented by machine learning, suggesting the need for further refinement.

Lee and colleagues address another aspect of privacy preservation based on the contention that unique trajectories of patients over time make it easier to reidentify patients.[2] They focus their attention on set-valued sequences that describe chronological medical conditions of patients with the goal of learning and synthesizing realistic sequences of EHR data. They developed the dual adversarial autoencoder (DAAE) that learns set-valued sequences of medical entities by combining a recurrent autoencoder with 2 generative adversarial networks. They evaluated the performance of DAAE for diagnostic codes in the context of predictive modeling and plausibility as well as privacy preservation using MIMIC-III and the University of Texas Physicians clinical database. Their findings support the adequacy of DAAE performance for predictive modeling and clinical plausibility. In addition, the differentially private optimization aspect of their approach enabled generation of synthetic sequences without increasing the privacy leakage of patient data.

Roberts et al describe the ongoing Text REtrieval Conference (TREC)-COVID information retrieval (IR) shared task whose goal is to galvanize the informatics community and provide the necessary data to help answer 6 important questions using the COVID-19 Open Research Dataset.[3] These are: (1) What are the appropriate IR modalities (ad hoc search, filtering, question-answering, etc) for this kind of event? (2) What are effective methods for customizing the search engine to the specific needs of the situation? (3) Can existing data be leveraged (eg, via machine learning) to improve the search engine? (4) Can event-specific training data be created fast enough to have an impact? (5) How does one quantitatively evaluate the search engine's performance? (6) How likely is it that different search engines have divergent enough performance to merit a quantitative comparison during a crisis? Participants are given about week from topic release (30 initial topics with 5 additional topics in each release) to result submission and can submit up to 1000 documents for each topic (eg, coronavirus hydroxy-chloroquine, coronavirus social distancing impact). Topics and data submissions are available at https://ir.nist.gov/covidSubmit/data.html. Performance is ranked per round. The processes and outcomes of the TREC-COVID shared task serve multiple purposes that influence discovery, care, and public health: (1) immediate support for researchers and clinicians fighting the pandemic; (2) development of a new IR evaluation process as the document collection, state of knowledge, and user interests evolve; and (3) a collection and approach to developing and implementing systems capable of satisfying information needs during pandemics.

Standardized vocabularies enable secondary use of EHR data for research and health information exchange. Dong and coauthors describe the development and evaluation of a rule-based tool called COVID-19 TestNorm that automatically normalizes local COVID-19 testing names to standard Logical Object Identifiers, Names, and Codes (LOINC).[5] Using 568 test names (454 for development and 114 for testing) collected from 8 healthcare systems, COVID-19 TestNorm achieved an accuracy of 97.4% on the test set. COVID-19 TestNorm is available as an open-source package for developers and as an online web application for end users (https://clamp.uth.edu/covid/loinc.php).

The Centers for Disease Control and Prevention (CDC) COVID-19 Information Management Repository was created to address the need for public health and healthcare stakeholders to easily obtain access to comprehensive and up-to-date information management resources including those aimed at improving interoperability.[5] Garcia et al provide an overview of 6 categories of COVID-19 resources in the Repository: (1) General (eg, LOINC Minimum Data Set for Public Health Emergency Operations Centers); (2) Emergency Medical Services (eg, National Emergency Medical Services Information System version 3 data dictionary); (3) Clinical Encounter (eg, SNOMED codes for COVID-19 patient encounters); (4) COVID-19 Public Health Reporting and Surveillance (eg, CDC COVID-19 Patient Impact and Hospital Capacity Module Form); (5) Laboratory Data Exchange and Laboratory Surveillance (eg, LOINC special use COVID-19 laboratory codes); (6) Geospatial Data Sets and Reference Sources (eg, CDC World COVID-19 Map). The repository is publicly available, distributed through CDC's Public Health Information Network Vocabulary Access and Distribution System (https://phinvads.cdc.gov/vads/SearchVocab.action).

The COVID-19 pandemic has certainly accelerated the application of biomedical and health informatics research to support discovery, care delivery, and health. At JAMIA, we remain particularly interested in innovative, scalable, and generalizable approaches that address important challenges to human health and health equity.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Carrell DS, Malin BA, Cronkite DJ, Aberdeen JS, *et al*. Resilience of clinical text de-identified with "hiding in plain sight" to hostile re-identification attacks by human readers. *J Am Med Inform Assoc* 2013; 20 (2): 342–8.
2. Lee D, Yu H, Jiang X, *et al*. Generating sequential electronic health records using dual adversarial autoencoder. *J Am Med Inform Assoc* 2020; 27 (9): 1411–19.
3. Roberts K, Alam T, Bedrick S, *et al*. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *J Am Med Inform Assoc* 2020; 27 (9): 1431–6.
4. Dong X, Li J, Soysal E, *et al*. COVID-19 TestNorm - A tool to normalize COVID-19 testing names to LOINC codes. *J Am Med Inform Assoc* 2020; 27 (9): 1437–42.
5. Garcia M, Lipskiy N, Tyson J, Watkins R, Esser ES, Kinley T. CDC COVID-19 information management: addressing national healthcare and public health needs for standardized data definitions and codified vocabulary for data exchange. *J Am Med Inform Assoc* 2020; 27 (9): 1476–87.