Print this Page for Your Records Close Window

Control/Tracking Number: 22-A-5373-AACR
Activity: Abstract Submission
Current Date/Time: 11/18/2021 9:08:08 PM

Developing a standardized framework for curating oncology datasets generated by manual abstraction and artificial intelligence

Author Block: *Benjamin M. M. Grant*, *Aein Zarrin*, *Luna Zhan*, *Rami Ajaj*, *Lina Darwish*, *Khaleeq Khan*, *Devalben Patel*, *Kaitlyn Chiasson*, *Karmugi Balaratnam*, *Maisha T. Chowdhury*, *Amir-Arsalan Sabouhanian*, *Joshua Herman*, *Preet Walia*, *Evan Strom*, *Catherine Brown*, *Miguel Garcia-Pardo*, *Sabine Schmid*, *Christopher Pettengell*, *Erin L. Stewart*, *Geoffrey Liu*. University Health Network, Toronto, ON, Canada, University of Toronto, Toronto, ON, Canada, Pentavere Research Group Inc., Toronto, ON, Canada, University of Toronto, Toronto, ON, Canada, Kantonsspital St. Gallen, Sankt Gallen, Switzerland

Abstract:

BACKGROUND: The widespread uptake of electronic health records (EHRs) has made the creation of custom, real-world datasets for research more feasible. As a result, multiple research datasets with overlapping populations are often generated, using different methodologies, and frequently siloed within and between research groups, limiting the scope of the data's use. Currently, there is no standard for collating and evaluating such data. Using existing lung oncology datasets, we developed an approach to determine optimal methods of combining and curating clinical data from different sources.

METHODS: Two separate study datasets containing data for lung cancer patients diagnosed and/or treated within Princess Margaret Cancer Centre (PM, Toronto) were investigated. Study 1 manually abstracted clinical data for 1,990 patients, first seen at PM between 2014-2016; Study 2 leveraged the artificial intelligence engine, DARWEN™, to extract clinical data directly from EHRs for 4,466 patients, diagnosed between 2014-2018. Each dataset was individually assessed for internal consistency before comparing the overlapping population (Test Group, n=1892) to identify, investigate, and resolve differences. Patterns of data extraction performance were evaluated to define optimal methods for combining datasets and informing future data collection. Herein, epidermal growth factor receptor (EGFR) mutation status is used as an illustrative example.

RESULTS: Study 1 and 2 had similar distributions of clinicodemographic data and frequency of EGFR mutations. The Test Group had 100% agreement for date of birth, and >99% agreement for sex, with all discrepancies resulting from human error in Study 1. The Test Group had a 98% agreement for EGFR positivity and 98-99% agreement for specific exon mutations. Of the 106 disagreements for specific mutations, 50% (n=53) were due to Study 1 human error. Study 2 prioritized specificity over sensitivity for biomarker extraction, resulting in more false negatives (25% of errors, n=26). As DARWEN™ only extracted EGFR data from pathology reports, 18% (n=19) of discrepancies were due to lack of access to relevant information captured elsewhere in patients' EHRs. Adjudicators could not resolve the remaining 7% of disagreements (n=8).

CONCLUSIONS: By comparing overlapping datasets, the strengths and weaknesses of each study design and extraction methodology were identified. This process demonstrated the effectiveness of artificial intelligence for extracting accurate patient-level clinicodemographic and mutation status data from EHRs, and the value of targeted manual chart review. Our approach provides a roadmap for leveraging existing clinical datasets to their fullest potential ,which is relevant across diverse data extraction methods and study designs.

Author Disclosure Information:

B.M.M. Grant: None. **A. Zarrin:** None. **L. Zhan:** None. **R. Ajaj:** None. **L. Darwish:** ; Pentavere Research Group Inc.. **K. Khan:** None. **D. Patel:** None. **K. Chiasson:** None. **K. Balaratnam:** None. **M.T. Chowdhury:** None. **A. Sabouhanian:** None. **J. Herman:** None. **P. Walia:** None. **E. Strom:** None. **C. Brown:** None. **M. Garcia-Pardo:** None. **S. Schmid:** None. **C. Pettengell:** ; Pentavere Research Group Inc.. ; Pentavere Research Group Inc. **E.L. Stewart:** ; Pentavere Research Group Inc.. ; Pentavere Research Group Inc. **G. Liu:** ; Honoraria; Abbvie. ; Abbvie. ; AstraZeneca. ; Honoraria; AstraZeneca. ; Honoraria; Bayer. ; Bayer. ; Bristol-Myers Squibb. ; Honoraria; Bristol-Myers Squibb. ; Honoraria; Merck. ; Merck. ; Novartis. ; Honoraria; Novartis. ; Honoraria; Pfizer. ; Pfizer. ; Roche Canada. ; Honoraria; Roche Canada. ; Honoraria; Takeda. ; Takeda.

Sponsor (Complete):
Category and Subclass (Complete): CL13-04 Other
Travel Support/Grants (Complete):
Is this study/trial supported in whole or in part by an AACR grant?: No
Is this study/trial sponsored in whole or in part by the pharmaceutical industry?: No

Poster Presentation Format (Complete): PRESENT IN PERSON. In addition to submitting an e-poster, I plan to present my poster in New Orleans.
Organ Site/Structures (Complete):
***Primary Organ Site:** Lung cancer: non-small cell
***Choose Chemical Structure Disclosure Option:**
NOT APPLICABLE. No compounds with defined chemical structures were used.
***Please explain reason for not disclosing (maximum 250 characters with spaces):** : NA
***Reference or patent application number** : NA

Keywords/Indexing (Complete): Lung cancer: non-small cell ; EGFR ; Databases ; Machine learning
Manuscript Publication (Complete):
Manuscript Publication Options: NO, I DO NOT EXPECT to submit a manuscript based upon my abstract at this time.