Cancer Informatics 2023: Data Sharing and Federating Learning Point Towards New Collaborative Opportunities

Jeremy L. Warner¹², Debra Patt³, Section Editors for the IMIA Yearbook Section on Cancer Informatics

¹ Professor, Department of Medicine, Brown University, Providence, RI, USA
² Director, Center for Clinical Cancer Informatics and Data Science, Legorreta Cancer Center, Brown University, Providence, RI, USA
³ Vice President, Texas Oncology, Austin, TX, USA

1 Introduction

Cancer informatics (CI) is a broad field with several fundamental goals: 1) organizing data in ways that are comprehensible and meaningful to clinicians, researchers, and patients; 2) using data to advance the treatment of cancer; and 3) manipulating data to yield new insights. In this edition of the Cancer Informatics section, we continue to focus on translational and clinical cancer informatics, with a special emphasis on data sharing in concordance with the 2023 Yearbook theme. As pointed out by Aneja, et al. [1] in the survey paper of the Cancer Informatics section of this IMIA Yearbook, “Despite growing enthusiasm surrounding the utility of clinical informatics to improve cancer outcomes, data availability remains a persistent bottleneck to progress. Difficulty combining data with protected health information often limits our ability to aggregate larger more representative datasets for analysis. Decentralized analytics, homomorphic encryption, and common data models represent promising solutions to improve cancer data sharing.” In order to overcome these challenges, technology solutions will need to scale beyond pilot and demonstration projects to national and international scales.

In 2023, the selection of papers in cancer informatics intends to illuminate the current progress of research with a focus on efforts to translate research towards immediate clinical applicability.

2 Paper Selection Method

One electronic database was searched, PubMed/MEDLINE. The search was performed in January 2023 to identify peer-reviewed journal articles published in 2022, in the English language, related to cancer informatics research. The following search was implemented:


This is identical to the search for papers in 2021, including that the MeSH terms related to computer-assisted radiotherapy planning were excluded due to previously observed high rates of false positives. This search yielded 1,952 results. Next, we excluded review articles, resulting in 1,826 articles for first-pass review. The titles of these articles were blindly screened for relevance, resulting in 109 articles that were reviewed in further depth. The titles of these articles were blindly screened for relevance, resulting in 109 articles that were reviewed in further depth. The abstract of each of the 109 was blindly reviewed and assigned as potential candidate (n=24), possible candidate (n=62), and non-candidate (n=23). Given that the number of potential candidates exceeded ten, these articles were further evaluated to select ten final candidates.
In accordance with the IMIA Yearbook selection process [2], the ten candidate best papers were evaluated by the two section editors, senior editors, and by additional external reviewers (at least four reviewers per paper). Three papers were finally selected as best papers (Table 1). A content summary of the selected best papers can be found in the appendix of this synopsis.

### 3 Conclusions and Outlook

The three selected best papers cover a variety of topics of important relevance to cancer informatics:

- **Pati et al.** [3] present what they report to be the largest federated learning study to date, involving data from 71 sites across 6 continents, to generate an automatic tumor boundary detector for the rare disease of glioblastoma. They demonstrate improvements in the delineation of surgically excisable tumor, and tumor extent, over a model that was trained on public data. This clinically relevant proof-of-principle demonstrates that federated learning can be used to address important and clinically relevant cancer topics.

- **Kuru et al.** [4] address the problem of trying to find synergistic drug combinations through the use of a deep learning framework. This work, entitled MatchMaker, seeks to overcome a serious bottleneck in the identification of new possibly efficacious combinations of drugs, which is highly relevant to the treatment of cancer. They report substantial improvements in correlation and mean squared error over the next best method.

- **Kondratieff et al.** [5] utilize electronic health record notes to develop clusters of topics using unsupervised topic modeling techniques. A two-stage modeling process built upon correlated topic modeling and structural topic modeling was able to identify clinically relevant topics in the notes of patients with breast cancer, including topical trends over time. This type of approach may surface unrecognized patient needs and may also enable proactive interventions for treatment-related and other toxicities.

The other candidate best papers cover the gamut of cancer informatics.

**Fang et al.** [6] tackle the challenge of distinguishing driver genes from passenger or neutral genes, a problem that has challenged the field for many years [7].

**Zhang et al.** [8] develop another method to approach the problem of identifying synergistic drug combinations, DCE-DForest. While similar in scope to Kuru et al. [4], this paper was felt to be comparatively less developed, leading to its honorable mention status.

- **Cancer Informatics**

<table>
<thead>
<tr>
<th>Paper</th>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
</table>

Finally, Rogier et al. [13] introduce OncoTox, an ontology designed to represent chemotherapy toxicities sourced from multiple sources such as electronic health record (EHR) questionnaires, semi-structured tables, and free text. Despite the increasing implementation of patient-reported outcomes, the EHR remains an invaluable source of toxicity information, which is a necessary component of understanding the patient with cancer’s journey, given that overall quality of life and many treatment discontinuation decisions are driven by toxicity.

### Acknowledgement

We would like to thank Kate Fultz Hollis, Lina Soualmia, and Adrien Ugon for their support and the reviewers for their participation in the selection process of the IMIA Yearbook.

### References

Cancer Informatics 2023: Data Sharing and Federating Learning Point Towards New Collaborative Opportunities

TCBB.2021.3086702.

Correspondence to:
Jeremy L. Warner MD, MS, FAMIA, FASCO (he/him/his)
Professor of Medicine at Brown University
Editor-in-Chief, JCO Clinical Cancer Informatics
E-mail: jeremy_warner@brown.edu
Appendix: Summary of Best Papers Selected for the 2023 Edition of the IMIA Yearbook, Section Cancer Informatics (CI)

Federated learning enables big data for rare cancer boundary detection

The authors present what they report to be the largest federated learning study to date, involving data from 71 sites across 6 continents, to generate an automatic tumor boundary detector for the rare disease of glioblastoma. They demonstrate improvements in the delineation of surgically excisable tumor, and tumor extent, over a model that was trained on public data. This clinically relevant proof-of-principle demonstrates that federated learning can be used to address important and clinically relevant cancer topics.

Kuru HI, Tastan O, Cicek AE
MatchMaker: A Deep Learning Framework for Drug Synergy Prediction

The authors address the problem of trying to find synergistic drug combinations through the use of a deep learning framework. This work seeks to overcome a serious bottleneck in the identification of new possibly efficacious combinations of drugs, which is highly relevant to the treatment of cancer. They report substantial improvements in correlation and mean squared error over the next best method.

Kondratieff KE, Brown JT, Barron M, Warner JL, Yin Z
Mining Medication Use Patterns from Clinical Notes for Breast Cancer Patients Through a Two-Stage Topic Modeling Approach
AMIA Annu Symp Proc 2022 May 23;2022:303-12

The authors utilize electronic health record notes to develop clusters of topics using unsupervised topic modeling techniques. A two-stage modeling process built upon correlated topic modeling and structural topic modeling was able to identify clinically relevant topics in the notes of patients with breast cancer, including topical trends over time. This type of approach may surface unrecognized patient needs and may also enable proactive interventions for treatment-related and other toxicities.