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

**Jeremy L. Warner[1,2], Debra Patt[3], 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

## Summary

**Objective**: To summarize significant research contributions on cancer informatics published in 2022.

**Methods**: An extensive search using PubMed/MEDLINE was conducted to identify the scientific contributions published in 2022 that address topics in cancer informatics. The selection process comprised three steps: (i) ten candidate best papers were first selected by the two section editors, (ii) external reviewers from internationally renowned research teams reviewed each candidate best paper, and (iii) the final selection of three best papers was conducted by the editorial board of the Yearbook.

**Results**: The three selected best papers demonstrate advances in federated learning, drug synergy prediction, and utilization of clinical note data.

**Conclusion**: Cancer informatics continues to mature as a subfield of biomedical informatics. Applications of informatics methods to data sharing and federated approaches are especially notable in 2022.

## Keywords

Neoplasms; informatics; health information technology; disparities

Yearb Med Inform 2023:111-4
http://dx.doi.org/10.1055/s-0043-1768744

## 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:

((“Neoplasms”[Mesh] OR “chemotherapy”) AND (“Informatics”[Mesh] OR “cancer informatics” OR “ontologies”) AND (hasabstract[text] AND (“2022/01/01”[PDAT]: “2022/12/31”[PDAT]) AND English[lang])) NOT (“Radiotherapy Planning, Computer-Assisted”[Mesh]) NOT (“Radiotherapy, Computer-Assisted”[Mesh])

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 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.

**Table 1** Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2023 in the section 'Cancer Informatics'. The articles are listed in reverse alphabetical order of the first author's surname.

| Section |
| --- |
| Cancer Informatics |
| ▪ Pati S, Baid U, Edwards B, Sheller M, Wang SH, Reina GA, et al. Federated learning enables big data for rare cancer boundary detection. Nat Commun 2022 Dec 5;13(1):7346. |
| ▪ Kuru HI, Tastan O, Cicek AE. MatchMaker: A Deep Learning Framework for Drug Synergy Prediction. IEEE/ACM Trans Comput Biol Bioinform 2022 Jul-Aug;19(4):2334-44. |
| ▪ 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. |

# 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. In a somewhat related vein, Kaczmarek *et al.* [9] utilize multi-omic data streams to classify cancer. This type of work seeks to change the histology-based paradigm that has been in place for well over 100 years, and is a welcome example of taking advantage of large dimensional data streams. Pu *et al.* [10] also take a multi-omic approach to the task of anti-cancer drug profiling through CancerOmicsNet, a graph neural network designed to predict the therapeutic effects of kinase inhibitors across various tumors.

Following the theme of synergy introduced by MatchMaker, Zeng *et al.* [11] developed RetriLite, an information retrieval and extraction framework that leverages natural language processing and a domain-specific knowledgebase to find potential signals for combination therapy in cancer. This is also one of several candidate papers to use transformer models, which have become quite notorious in the past year with the introduction of commercial systems such as Chat-GPT and Google Bard, following on earlier efforts such as BERT and RoBERTa. Yu *et al.* [12] applied BERT and RoBERTa to the important task of extracting social and behavioral determinants of health from clinical narratives, with fairly good success (strict F1=0.8791).

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

1. Aneja S, Avesta A, Xu H, Machado LO. Clinical Informatics Approaches to Facilitate Cancer Data Sharing. Yearb Med Inform 2023 Jul 6. doi: 10.1055/s-0043-1768721.
2. Lamy J-B, Séroussi B, Griffon N, Kerdelhué G, Jaulent M-C, Bouaud J. Toward a formalization of the process to select IMIA Yearbook best papers. Methods Inf Med 2015;54(2):135–44. doi: 10.3414/ME14-01-0031.
3. Pati S, Baid U, Edwards B, Sheller M, Wang SH, Reina GA, et al. Federated learning enables big data for rare cancer boundary detection. Nat Commun 2022 Dec 5;13(1):7346. doi: 10.1038/s41467-022-33407-5.
4. Kuru HI, Tastan O, Cicek AE. MatchMaker: A Deep Learning Framework for Drug Synergy Prediction. IEEE/ACM Trans Comput Biol Bioinform 2022 Jul-Aug;19(4):2334-44. doi: 10.1109/

TCBB.2021.3086702.
5. 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.
6. Fang H, Zhang Z, Zhou Y, Jin L, Yang Y. A greedy approach for mutual exclusivity analysis in cancer study. Biostatistics 2022 Jul 18;23(3):910-25. doi: 10.1093/biostatistics/kxab004.
7. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science 2013 Mar 29;339(6127):1546-58. doi: 10.1126/science.1235122.
8. Zhang W, Xue Z, Li Z, Yin H. DCE-DForest: A Deep Forest Model for the Prediction of Anticancer Drug Combination Effects. Comput Math Methods Med 2022 Jun 9;2022:8693746. doi: 10.1155/2022/8693746.
9. Kaczmarek E, Jamzad A, Imtiaz T, Nanayakkara J, Renwick N, Mousavi P. Multi-Omic Graph Transformers for Cancer Classification and Interpretation. Pac Symp Biocomput 2022;27:373-84.
10. Pu L, Singha M, Ramanujam J, Brylinski M. CancerOmicsNet: a multi-omics network-based approach to anti-cancer drug profiling. Oncotarget 2022 May 19;13:695-706. doi: 10.18632/oncotarget.
11. Zeng J, Cruz-Pico CX, Saridogan T, Abu Shufean M, Kahle M, Yang D, et al. Natural Language Processing-Assisted Literature Retrieval and Analysis for Combination Therapy in Cancer. JCO Clin Cancer Inform 2022 Jan;6:e2100109. doi: 10.1200/CCI.21.00109.
12. Yu Z, Yang X, Dang C, Wu S, Adekkanattu P, Pathak J, George TJ, et al. A Study of Social and Behavioral Determinants of Health in Lung Cancer Patients Using Transformers-based Natural Language Processing Models. AMIA Annu Symp Proc 2022 Feb 21;2021:1225-33.
13. Rogier A, Coulet A, Rance B. Using an Ontological Representation of Chemotherapy Toxicities for Guiding Information Extraction and Integration from EHRs. Stud Health Technol Inform 2022 Jun 6;290:91-5. doi: 10.3233/SHTI220038.

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)

### Pati S, Baid U, Edwards B, Sheller M, Wang SH, Reina GA, et al

### Federated learning enables big data for rare cancer boundary detection

### Nat Commun 2022 Dec 5;13(1):7346. doi: 10.1038/s41467-022-33407-5

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

### IEEE/ACM Trans Comput Biol Bioinform 2022 Jul-Aug;19(4):2334-2344. doi: 10.1109/TCBB.2021.3086702

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.