Article published online: 2022-12-04

© 2022 IMIA and Georg Thieme Verlag KG

Cancer Informatics 2022: Real-World Data Yields Important Insights into the Conduct of Clinical Trials and Registries

Jeremy L. Warner[1,2], **Michael K. Rooney**[3], **Debra Patt**[1,4]

[1] Section Editors for the IMIA Yearbook Section on Cancer Informatics
[2] Associate Professor, Departments of Medicine and Biomedical Informatics, Vanderbilt University, Nashville, TN, USA
[3] Radiation Oncology Resident, MD Anderson Cancer Center, Houston, TX, USA
[4] Vice President, Texas Oncology, Austin, TX, USA

Summary

Objective: To summarize significant research contributions on cancer informatics published in 2021.

Methods: An extensive search using PubMed/MEDLINE and Altmetric scores was conducted to identify the scientific contributions published in 2021 that address topics in cancer informatics. The selection process comprised three steps: (i) 15 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 two best papers was conducted by the editorial board of the IMIA Yearbook.

Results: The two selected best papers demonstrate some of the promises and shortcomings of real-world data.

Conclusion: Cancer informatics is a maturing subfield of biomedical informatics. Applications of informatics methods to real-world data are especially notable in 2021.

Keywords

Neoplasms; informatics; health information technology; disparities

Yearb Med Inform 2022:131-5
http://dx.doi.org/10.1055/s-0042-1742521

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 fourth year of the Cancer Informatics section (there was no CI section in 2021, due to impacts of the COVID-19 pandemic), we continue to focus on translational and clinical cancer informatics, with a special emphasis on disparities in concordance with the 2022 Yearbook theme. As pointed out by Chaunzwa, *et al.*, [1] in the survey paper of the Cancer Informatics section of this IMIA Yearbook, "*As informatics tools become integrated into clinical decision-making, attention will need to be paid to ensure that algorithmic bias does not amplify existing disparities. In our increasingly interconnected medical systems, clinical informatics is poised to untap the full potential of multi-platform health data to address cancer disparities*". In order to overcome these challenges, technology solutions cannot be considered in a vacuum, even those with very high performance.

In 2022, 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 2022 to identify peer-reviewed journal articles published in 2021, 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 (“2021/01/01”[PDAT] : “2021/12/31”[PDAT]) AND English[lang])) NOT (“Radiotherapy Planning, Computer-Assisted”[MeSH]) NOT (“Radiotherapy, Computer-Assisted”[MeSH]).

This search includes several exclusion terms related to computer-assisted radiotherapy planning, to avoid previously observed high rates of false positives. This search yielded 3,863 results, an infeasibly large number of titles to review. Therefore, we implemented a new filtering step using Altmetric Attention scores (hereafter, “Altmetrics”) [2]. We assessed the Altmetrics for all articles as of January 18, 2022, using the rAltmetric R package [3]. Altmetrics ranged from 0/incalculable to 1730.6 (Figure 1). A threshold of Altmetric score of 20 was applied, resulting in 200 candidates for additional review. The titles of these 200 articles were manually screened by one of the two

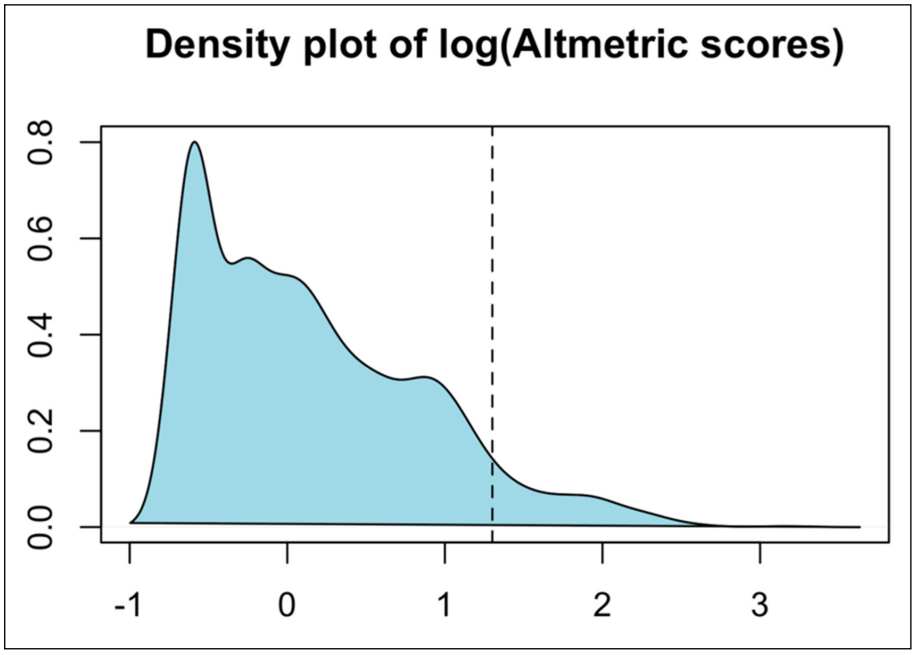

Fig. 1 The density distribution of Altmetric scores for the identified articles shows that a relatively small number have very high Altmetrics. The threshold (vertical dashed line), which was chosen prior to examination of the data, includes this distribution.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2022 in the section 'Cancer Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section Cancer Informatics
▪ Liu R, Rizzo S, Whipple S, Pal N, Pineda AL, Lu M, Arnieri B, Lu Y, Capra W, Copping R, Zou J. Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature 2021 Apr;592(7855):629-33.
▪ Yang DX, Khera R, Miccio JA, Jairam V, Chang E, Yu JB, Park HS, Krumholz HM, Aneja S. Prevalence of Missing Data in the National Cancer Database and Association With Overall Survival. JAMA Netw Open 2021 Mar 1;4(3):e211793.

section editors, and the abstracts of 59 of these were manually reviewed by the same editor in order to arrive at a candidate list of 15 papers. These articles were selected as final candidates.

In accordance with the IMIA Yearbook selection process [4], the 15 candidate best papers were evaluated by the two section editors and by additional external reviewers (at least four reviewers per paper). Two 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 two selected best papers deal with complementary aspects of real-world databases and registries, which continue to increase in popularity for population-level research as well as regulatory decision-making.

Liu, *et al.,* [5] describe a computational framework called Trial Pathfinder that was developed and evaluated on a series of non-small cell lung cancer (NSCLC) trials to determine whether changing trial eligibility criteria would change the results. The simulations were conducted using a large real-world data source of 61,094 patients with advanced NSCLC sourced from the Flatiron Health database, a commercial database of curated electronic medical record data. The authors determined that relaxing eligibility criteria would result in small changes to hazard ratios for critical outcomes such as overall survival, while potentially expanding the pool of eligible patients by at least two-fold. This has major implications for the future design of cancer clinical trials and is especially pertinent to the Yearbook theme of disparities. For example, women, older adults, and racial/ethnic minorities have been relatively excluded from cancer clinical trials, and this trend appears to be worsening over time [6].

Yang, *et al*., [7] assessed the prevalence of missing data in a very large registry of patients with cancer, and whether missingness in itself was prognostic. The registry evaluated was the National Cancer Database (NCDB) maintained by the Commission on Cancer. While this is not a true population-based registry, it is nevertheless a very large voluntary registry that has been extensively used in outcomes research [8]. They found substantial missingness, e.g., with 71% of patients with NSCLC (n=851,295) missing data for variables of interest. When compared to patients with complete case data, they found a statistically significant difference in 2-year overall survival, consistent across three important cancer subtypes (lung, breast, prostate). This study demonstrates the importance of metadata characteristics in the conduct and evaluation of real-world data registry studies. The authors note that "Records with missing data were more prevalent among Black patients and patients from other racial and ethnic minority groups, which may reflect long-standing disparities in access to health care and cancer treatment". Whether or not it is a proxy, data missingness must be considered within the larger framework of disparities.

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

In keeping with the Yearbook theme, Awasthi, *et al.*, [9] discovered distinct immune-oncologic pathways in African American men with prostate cancer. Race-specific differences in gene expression were found to have implications for treatment approaches,

which should be evaluated in prospective fashion. These types of studies sit at a challenging juncture between race, a social construct, and ancestry, a biologic measure.

Warnat-Herresthal, *et al.,* [10] described a decentralized machine-learning approach for privacy-preserving studies. While one of their use cases (leukemia) was in the cancer domain, this study was felt to be too generic for the CI section.

Several highly meritorious papers were felt to have too much overlap with other sections of the Yearbook, in particular the Bioinformatics and Translational Informatics section. Bagaev, *et al.,* [11] found that conserved pan-cancer microenvironment subtypes were predictive of immunotherapy response, which could be particularly relevant for a subset of difficult-to-treat cancers. Cheng, *et al.,* [12] characterized protein-protein interactions in nearly 11,000 tumor exomes and demonstrate a correlation with patient survival and resistance to drugs. Dentro, *et al.,* [13] characterized genetic intra-tumor heterogeneity across more than 2,500 human cancer genomes. Hu, *et al.,* [14] describe SpaGCN, a graph convolutional network approach using multimodal data: histology, gene expression, and spatial location. Scott, *et al.,* [15] further developed the concept of genomic-adjusted radiation dose using a cohort-based pooled analysis.

Cantini, *et al.,* [16] undertook a systemic benchmarking evaluation of nine joint Dimensionality Reduction (jDR) methods. They found that one method (intNMF) excels in clustering, whereas another (MCIA) has good all-around performance. Given the panoply of bioinformatics algorithms available, these types of benchmarking efforts are most welcome.

Absolom, *et al.,* [17] reported on one of the first phase three randomized clinical trials examining the effects of an eHealth intervention during chemotherapy. They found that the Electronic patient self-Reporting of Adverse-events: Patient Information and aDvice (eRAPID) system, compared to usual care, improved physical well-being at 6 and 12 weeks, although there was no difference at 18 weeks (primary end point). We hope this is the first of many such rigorous evaluations.

Gould, *et al.,* [18] describe an approach to identify early lung cancer using routine laboratory and clinical data. They report an AUC of 0.86 for identifying NSCLC up to one year prior to clinical diagnosis, outperforming a previously validated prediction model, the 2012 Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial risk model.

Zhou, *et al.,* [19] conducted an interesting study on the effect of adversarial images within the domain of breast cancer mammography. They found that human radiologists could identify the adversarial images at a generally higher rate of success than a deep learning-based computer-aided diagnosis system, illustrating the need for ongoing work in this area.

Do, *et al.,* [20] use natural language processing to discover sites of metastatic disease in radiology reports collected over a 10-year period. This encouraging paper was ultimately not selected as a best paper mostly due to the concern for generalizability, since the analyzed corpus consisted of structured radiology reports that are likely unique to the study institution.

Finally, Mendiratta, *et al.,* [21] estimate the mutation frequencies for cancer genes across the US population. While this is a simulation study, it nevertheless suggests that there could be important systemic differences in gene mutation frequencies than are reported in large public cancer sequencing registries.

Acknowledgement

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

References

1. Chaunzwa TL, Quiles del Rey M, Bitterman DS. Clinical Informatics Approaches to Understand and Address Cancer Disparities. Yearb Med Inform 2022:121-30.
2. Ram K. rAltmetric: Retrieves altmetrics data for any published paper from altmetrics.com. R package version 0.7. Available from: http://CRAN.R-project.org/package=rAltmetric
3. Galligan F, Dyas-Correia S. Altmetrics: Rethinking the Way We Measure. Serials Review 2013 Mar;39(1):56–61.
4. Lamy JB, Séroussi B, Griffon N, Kerdelhué G, Jaulent MC, Bouaud J. Toward a formalization of the process to select IMIA Yearbook best papers. Methods Inf Med 2015;54(2):135-44.
5. Liu R, Rizzo S, Whipple S, Pal N, Pineda AL, Lu M, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature 2021 Apr;592(7855):629-33.
6. Riaz IB, Islam M, Khan AM, Naqvi SAA, Siddiqi R, Khakwani KZR, et al. Disparities in Representation of Women, Older Adults, and Racial/Ethnic Minorities in Immune Checkpoint Inhibitor Trials. Am J Med 2022 Apr 25:S0002-9343(22)00328-X.
7. Yang DX, Khera R, Miccio JA, Jairam V, Chang E, Yu JB, et al. Prevalence of Missing Data in the National Cancer Database and Association With Overall Survival. JAMA Netw Open 2021 Mar 1;4(3):e211793.
8. Boffa DJ, Rosen JE, Mallin K, Loomis A, Gay G, Palis B, et al. Using the National Cancer Database for Outcomes Research: A Review. JAMA Oncol 2017 Dec 1;3(12):1722-8.
9. Awasthi S, Berglund A, Abraham-Miranda J, Rounbehler RJ, Kensler K, Serna A, et al. Comparative Genomics Reveals Distinct Immune-oncologic Pathways in African American Men with Prostate Cancer. Clin Cancer Res 2021 Jan 1;27(1):320-9.
10. Warnat-Herresthal S, Schultze H, Shastry KL, Manamohan S, Mukherjee S, Garg V, et al. Swarm Learning for decentralized and confidential clinical machine learning. Nature 2021 Jun;594(7862):265-70.
11. Bagaev A, Kotlov N, Nomie K, Svekolkin V, Gafurov A, Isaeva O, et al. Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. Cancer Cell 2021 Jun 14;39(6):845-65.e7.
12. Cheng F, Zhao J, Wang Y, Lu W, Liu Z, Zhou Y, et al. Comprehensive characterization of protein-protein interactions perturbed by disease mutations. Nat Genet 2021 Mar;53(3):342-53.
13. Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, et al; PCAWG Evolution and Heterogeneity Working Group and the PCAWG Consortium. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. Cell 2021 Apr 15;184(8):2239-54.e39.
14. Hu J, Li X, Coleman K, Schroeder A, Ma N, Irwin DJ, et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat Methods 2021 Nov;18(11):1342-51.
15. Scott JG, Sedor G, Ellsworth P, Scarborough JA, Ahmed KA, Oliver DE, et al. Pan-cancer prediction of radiotherapy benefit using genomic-adjusted radiation dose (GARD): a cohort-based pooled analysis. Lancet Oncol 2021 Sep;22(9):1221-9.
16. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat Commun 2021 Jan 5;12(1):124.

17. Absolom K, Warrington L, Hudson E, Hewison J, Morris C, Holch P, et al. Phase III Randomized Controlled Trial of eRAPID: eHealth Intervention During Chemotherapy. J Clin Oncol 2021 Mar 1;39(7):734-47.
18. Gould MK, Huang BZ, Tammemagi MC, Kinar Y, Shiff R. Machine Learning for Early Lung Cancer Identification Using Routine Clinical and Laboratory Data. Am J Respir Crit Care Med 2021 Aug 15;204(4):445-53.
19. Zhou Q, Zuley M, Guo Y, Yang L, Nair B, Vargo A, et al. A machine and human reader study on AI diagnosis model safety under attacks of adversarial images. Nat Commun 2021 Dec 14;12(1):7281.
20. Do RKG, Lupton K, Causa Andrieu PI, Luthra A, Taya M, Batch K, et al. Patterns of Metastatic Disease in Patients with Cancer Derived from Natural Language Processing of Structured CT Radiology Reports over a 10-year Period. Radiology 2021 Oct;301(1):115-22.
21. Mendiratta G, Ke E, Aziz M, Liarakos D, Tong M, Stites EC. Cancer gene mutation frequencies for the U.S. population. Nat Commun 2021 Oct 13;12(1):5961.

Correspondence to:
Jeremy L. Warner MD, MS
Associate Professor
of Medicine and Biomedical Informatics
Vanderbilt University Medical Center
2220 Pierce Avenue, 777 PRB
Nashville, TN 37232-6307
USA
E-mail: jeremy.warner@brown.edu