Published online: 16.08.2019

Cancer Informatics in 2018: The Mysteries of the Cancer Genome Continue to Unravel, Deep Learning Approaches the Clinic, and Passive Data Collection Demonstrates Utility

Jeremy L. Warner[1], Debra Patt[2], Section Editors for the IMIA Yearbook Section on Cancer Informatics

[1] Associate Professor, Departments of Medicine and Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

[2] Vice President, Texas Oncology, Austin, TX, USA

Summary

Objective: To summarize significant research contributions on cancer informatics published in 2018.

Methods: An extensive search using PubMed/Medline, Google Scholar, and manual review was conducted to identify the scientific contributions published in 2018 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 four best papers was conducted by the editorial board of the International Medical Informatics Association (IMIA) Yearbook.

Results: The four selected best papers present studies addressing many facets of cancer informatics, with immediate applicability in the translational and clinical domains.

Conclusion: Cancer informatics is a broad and vigorous subfield of biomedical informatics. Progress in cancer genomics, artificial intelligence, and passively collected data is especially notable in 2018.

Keywords

Neoplasms, informatics, health information technology, genomics, ontologies

Yearb Med Inform 2019:236-9
http://dx.doi.org/10.1055/s-0039-1677931

Introduction

The field of cancer informatics intends to take full advantage of the many data streams generated in the course of cancer prevention, diagnosis, care, and survivorship with several fundamental goals: 1) organizing the data in ways that are comprehensible and meaningful to clinicians, researchers, and patients; 2) using the data to advance the treatment of cancer; 3) bringing new data streams, such as person-generated data, into the mix; and 4) manipulating the data, such as through visualization, to yield new insights. In this second year of the Cancer Informatics section, we continue to focus on translational and clinical cancer informatics. While there is no survey paper this year, progress continues to be rapid, most notably in the area of cancer genomics.

In 2019, 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.

Paper Selection Method

Two electronic databases were searched, PubMed/MEDLINE and Google Scholar. Searches were performed in January 2019 to identify peer-reviewed journal articles. A PubMed search using the MeSH terms "Neoplasms" and "Medical Informatics Applications" for papers in English language published between Oct 31 2017 and Dec 31 2018 returned too many results for practical review (3,479 results). Similarly, a Google Scholar search for "cancer informatics" limited to 2017 & 2018 returned 27,700 results. Thus, we performed searches of PubMed-indexed well-known informatics journals and proceedings (i.e., *JAMIA*, *Applied Clinical Informatics, Bioinformatics, Journal of Biomedical Informatics, Journal of Medical Internet Research, Pacific Symposium on Biocomputing*) using the search phrase ("J Am Med Inform Assoc"[Journal]) AND neoplasms[MeSH Major Topic] (as shown, for JAMIA). Additionally, the contents of the journal *JCO Clinical Cancer Informatics*, which was not PubMed-indexed in 2018, were hand-searched. For candidate articles that were PubMed-indexed, we also searched for additional relevant articles using PubMed's "Similar articles" service. We also searched the proceedings of AMIA 2018 and the 2018 AMIA Joint Summits.

One of the two section editors performed the searches. Given the vast results, we focused on identifying articles with translational or clinical applications, as opposed to papers describing more fundamental bioinformatics methodologies. Then, the two section editors undertook independently the initial screening of titles and abstracts to identify papers relevant to the field of interest. Both section editors classified the papers into three categories: definitely

include, possibly include, or exclude. They then reviewed in detail the possibly include full-text articles to finally reach a mutual list of 15 candidate best papers. Papers were considered according to their originality, innovativeness, scientific and/or practical impact, and scientific quality.

In accordance with the IMIA Yearbook selection process [1], the 15 candidate best papers were evaluated by the two section editors and by additional external reviewers (at least four reviewers per paper). Four 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 2019 in the section 'Cancer Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Cancer Informatics
▪ Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, Ng PK, Jeong KJ, Cao S, Wang Z, Gao J, Gao Q, Wang F, Liu EM, Mularoni L, Rubio-Perez C, Nagarajan N, Cortés-Ciriano I, Zhou DC, Liang WW, Hess JM, Yellapantula VD, Tamborero D, Gonzalez-Perez A, Suphavilai C, Ko JY, Khurana E, Park PJ, Van Allen EM, Liang H; MC3 Working Group; Cancer Genome Atlas Research Network, Lawrence MS, Godzik A, Lopez-Bigas N, Stuart J, Wheeler D, Getz G, Chen K, Lazar AJ, Mills GB, Karchin R, Ding L. Comprehensive characterization of cancer driver genes and mutations. Cell 2018 Apr 5;173(2):371-385.e18.
▪ Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, Bussink J, Gillies RJ, Mak RH, Aerts HJWL. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med 2018 Nov 30;15(11):e1002711.
▪ Low CA, Dey AK, Ferreira D, Kamarck T, Sun W, Bae S, Doryab A. Estimation of symptom severity during chemotherapy from passively sensed data: Exploratory study. J Med Internet Res 2017 Dec 19;19(12):e420.
▪ Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, Tusquets I, Albanell J, Rodon J, Tabernero J, de Torres C, Dienstmann R, Gonzalez-Perez A, Lopez-Bigas N. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med 2018 Mar 28;10(1):25.

Conclusions and Outlook

The four selected best papers are representative of three distinct subdomains of cancer informatics: cancer genomics; deep learning applications; and passive data collection.

Bailey, *et al.,* [2] described a very substantial international effort to catalogue a list of cancer driver genes and their mutations. Generally speaking, somatic mutations observed in cancers are felt to either be *drivers* of the cancer or passive *passengers*; most drug development is focused on disabling drivers. Despite much work in this area, automated algorithms often do not agree on candidate driver genes and mutations, requiring expert manual curation. This broad application of 26 bioinformatic software tools to 10,000 TCGA tumor samples (representing 33 cancer types) is the most comprehensive discovery of cancer driver genes and mutations to date. The data generated lay the groundwork for years of basic, translational, and clinical efforts.

Hosny, *et al.,* [3] conducted a multi-site retrospective study of lung cancer prognostication using radiomics. Despite being the deadliest cancer, there are scant prognostic tools to determine lung cancer prognosis outside of the traditional anatomic staging systems. This study was an integrative analysis on seven independent radiographic datasets across five institutions, using a 3D convolution neural network. The authors successfully predicted survival for lung cancer patients with an AUC of 0.70. While the performance is only modest, radiography is standard of care for lung cancer and this study has immediate clinical applicability for a very common and highly lethal disease. Over the coming years, we expect that such approaches will become more comprehensive and more accurate, such as the recently published [4] approach to lung cancer screening by the Google Artificial Intelligence (AI) group.

Low, *et al.,* [5] undertook a prospective pilot study of passively collected data from patients undergoing chemotherapy treatment. The aim of this study was to explore whether passively sensed mobile phone and Fitbit® data could be used to estimate daily symptom burden during chemotherapy. Passively collected data matched patient-reported symptom burden with an accuracy of 88%. This was a small study (n=14), but proof of concept that mobile phone accelerometer and usage and Fitbit®-assessed activity and sleep were related to daily symptom burden during chemotherapy. These findings highlight opportunities for long-term monitoring of cancer patients during chemotherapy, including the possibility of obviating invasive and tedious surveys. The interested reader is also referred to the recent American Society of Clinical Oncology educational book chapter [6] on the topic.

Tamborero, *et al.,* [7] presented Cancer Genome Interpreter (https://www.cancergenomeinterpreter.org/home), a software tool that streamlines and automates the process of identifying and annotating variants. The process is critical to interpretation and contextualization of tumor sequencing results, in particular calling out those that may have clinical actionability. The tool accepts several data formats and provides a user-friendly output. Also described is a new knowledge base of 5,314 validated mutations (the Catalog of Validated Oncogenic Mutations). As a proof of concept, 72% of AACR Project GENIE [8] tumors (~17k) have at least one biomarker of drug response in the system. This proportion is much higher than what has been reported in older studies of genomically-informed treatment decisions, suggesting that the match between mutation and drug continues to improve.

The other candidate best papers are in the same line with innovative and/or effective cancer informatics approaches.

Two of the papers are in the general domain of data extraction and data mining. Baker, *et al.,* [9] described the Cancer Hallmarks Analytics Tool (CHAT), which extracts cancer-relevant literature from Pubmed. Gianni, *et al.,* [10] described an effort that seeks to find new combination

treatments for cancer using genomics and clinical datasets. These complementary approaches of literature search and clinical data mining should become increasingly intertwined in the future.

Three of the candidate best papers [11–13] are in the cancer genomics domain, similar to Bailey, *et al.,* [2] and Tamborero, *et al*., [7]. Bertrand, *et al.,* [11] introduced ConsensusDriver, an algorithmic approach to adjudicating the discrepancies in driver mutation ascertainment mentioned above. Sun, *et al.,* [12] tackled an important related problem – determining the difference between germline and somatic mutations in the absence of matched normal tissue. Piñeiro-Yáñez, *et al.,* [13] developed PanDrugs, a method that aims to prioritize drugs by genomic findings in the cancer specimen.

Four of the candidate best papers [4–17] are concerned with knowledge management and ontologies. Cario, *et al.,* [14] and Warner, *et al.,* [15] described Orchid and SMART Cancer Navigator respectively, both of which are frameworks for the management of the knowledge needed to practice precision cancer medicine. Malty, *et al.,* [16] described an OWL-based ontology to represent chemotherapy regimens. Pecora, *et al.,* [17] described a barcode-like approach to classification of cancer.

Finally, several additional candidate papers described the uses of information technology for patient and caregiver engagement. Gupta, *et al.,* [18] carried out a feasibility study of using physical activity monitors as a surrogate for clinician-ascertained performance status. Gustafson, *et al.,* [19] conducted two sizeable randomized clinical trials examining the utility of a caregiver e-alert system to reduce patients' distress from symptoms.

Acknowledgement

We would like to thank Brigitte Seroussi for her support and the reviewers for their participation in the selection process of the IMIA Yearbook.

References

1. 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.
2. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. Cell 2018;173(2):371-385.e18.
3. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med 2018;15(11):e1002711.
4. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019 2019 Jun;25(6):954-61.
5. Low CA, Dey AK, Ferreira D, Kamarck T, Sun W, Bae S, et al. Estimation of symptom severity during chemotherapy from passively sensed data: Exploratory study. J Med Internet Res 2017;19(12):e420.
6. Liao Y, Thompson C, Peterson S, Mandrola J, Beg MS. The future of wearable technologies and remote monitoring in health care. Am Soc Clin Oncol Ed Book 2019 May 1;(39):115–21.
7. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med 2018;10(1):25.
8. AACR Project GENIE Consortium. AACR Project GENIE: Powering precision medicine through an international consortium. Cancer Discov 2017 Aug;7(8):818–31.
9. Baker S, Ali I, Silins I, Pyysalo S, Guo Y, Högberg J, et al. Cancer Hallmarks Analytics Tool (CHAT): a text mining approach to organize and evaluate scientific literature on cancer. Bioinformatics 2017 Dec 15;33(24):3973–81.
10. Gianni M, Qin Y, Wenes G, Bandstra B, Conley AP, Subbiah V, et al. High-throughput architecture for discovering combination cancer therapeutics. JCO Clin Cancer Inform 2018;2:1–12.
11. Bertrand D, Drissler S, Chia BK, Koh JY, Li C, Suphavilai C, et al. ConsensusDriver improves upon individual algorithms for predicting driver alterations in different cancer types and individual patients. Cancer Res 2018;78(1):290–301.
12. Sun JX, He Y, Sanford E, Montesion M, Frampton GM, Vignot S, et al. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. PLoS Comput Biol 2018;14(2):e1005965.
13. Piñeiro-Yáñez E, Reboiro-Jato M, Gómez-López G, Perales-Patón J, Troulé K, Rodríguez JM, et al. PanDrugs: a novel method to prioritize anticancer drug treatments according to individual genomic data. Genome Med 2018;10(1):41.
14. Cario CL, Witte JS. Orchid: a novel management, annotation and machine learning framework for analyzing cancer mutations. Bioinformatics 2018;34(6):936–42.
15. Warner JL, Prasad I, Bennett M, Arniella M, Beeghly-Fadiel A, Mandl KD, et al. SMART Cancer Navigator: A framework for implementing ASCO workshop recommendations to enable precision cancer medicine. JCO Precis Oncol 2018 May 1;(2):1–14.
16. Malty AM, Jain SK, Yang PC, Harvey K, Warner JL. Computerized approach to creating a systematic ontology of hematology/oncology regimens. JCO Clin Cancer Inform 2018 May 1;(2):1–11.
17. Pecora AL, Norden AD, Hervey J, Schultz EV, Gallucci TL, Rushforth E, et al. Development of a precise, clinically relevant, digital classification schema for cancer. JCO Clin Cancer Inform 2018;2:1–10.
18. Gupta A, Stewart T, Bhulani N, Dong Y, Rahimi Z, Crane K, et al. Feasibility of wearable physical activity monitors in patients with cancer. JCO Clin Cancer Inform 2018;2:1–10.
19. Gustafson DH, DuBenske LL, Atwood AK, Chih M-Y, Johnson RA, McTavish F, et al. Reducing symptom distress in patients with advanced cancer using an e-alert system for caregivers: Pooled analysis of two randomized clinical trials. J Med Internet Res 2017 14;19(11):e354.

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@vumc.org

Appendix: Summary of Best Papers Selected for the 2019 Edition of the IMIA Yearbook, Section Cancer Informatics

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, Ng PK, Jeong KJ, Cao S, Wang Z, Gao J, Gao Q, Wang F, Liu EM, Mularoni L, Rubio-Perez C, Nagarajan N, Cortés-Ciriano I, Zhou DC, Liang WW, Hess JM, Yellapantula VD, Tamborero D, Gonzalez-Perez A, Suphavilai C, Ko JY, Khurana E, Park PJ, Van Allen EM, Liang H; MC3 Working Group; Cancer Genome Atlas Research Network, Lawrence MS, Godzik A, Lopez-Bigas N, Stuart J, Wheeler D, Getz G, Chen K, Lazar AJ, Mills GB, Karchin R, Ding L

Comprehensive characterization of cancer driver genes and mutations

Cell 2018 Apr 5;173(2):371-385.e18

Understanding which genes and which gene mutations are cancer drivers is an essential first step towards contemplating ways to disable the cancer machinery through pharmacologic intervention. Generally speaking, somatic mutations observed in cancers are felt to either be *drivers* of the cancer or passive *passengers*; most drug development is focused on disabling drivers. Despite much work in this area, automated algorithms often do not agree on candidate driver genes and mutations, requiring expert manual curation. This broad application of 26 bioinformatic software tools to 10,000 TCGA tumor samples (representing 33 cancer types) is the most comprehensive discovery of cancer driver genes and mutations to date. The data generated lay the groundwork for years of basic, translational, and clinical efforts. All data generated are publicly available.

Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, Bussink J, Gillies RJ, Mak RH, Aerts HJWL

Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study

PLoS Med 2018 Nov 30;15(11):e1002711

Despite several years of anticipation, artificial intelligence methods such as deep learning have yet to enter the clinical cancer setting. In general, findings based on a single institution retrospective study must be replicated across institutions before prospective trials can be considered. Hosny *et al.,* have met the second mark through their multi-site retrospective study of lung cancer prognostication using radiomics. Despite being the deadliest cancer, there are scant prognostic tools to determine lung cancer prognosis outside of the traditional anatomic staging systems. This study was an integrative analysis on seven independent radiographic datasets across five institutions, using a 3D convolution neural network. The authors successfully predicted survival for lung cancer patients with an AUC of 0.70. While the performance is only modest, radiography is standard of care for lung cancer and this study has immediate clinical applicability for a very common and highly lethal disease. Over the coming years, we expect that such approaches will become more comprehensive and accurate and will be tested in the prospective setting.

Low CA, Dey AK, Ferreira D, Kamarck T, Sun W, Bae S, Doryab A

Estimation of symptom severity during chemotherapy from passively sensed data: Exploratory study

J Med Internet Res 2017 Dec 19;19(12):e420

With smart phones and other wearable devices now nearly ubiquitous, it is natural to wonder if they can be utilized as health care collection tools. In particular, can passive data collection yield insights similar to those collected directly from patients? Low, *et al.,* undertook a prospective pilot study of passively collected data from patients undergoing chemotherapy treatment. The aim of this study was to explore whether passively sensed mobile phone and Fitbit® data could be used to estimate daily symptom burden during chemotherapy. Passively collected data matched patient-reported symptom burden with an accuracy of 88%. This was a small study (n=14), but proof of concept that mobile phone accelerometer and usage and Fitbit®-assessed activity and sleep were related to daily symptom burden during chemotherapy. These findings highlight opportunities for long-term monitoring of cancer patients during chemotherapy, including the possibility of obviating invasive and tedious surveys.

Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, Tusquets I, Albanell J, Rodon J, Tabernero J, de Torres C, Dienstmann R, Gonzalez-Perez A, Lopez-Bigas N

Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations

Genome Med 2018 Mar 28;10(1):25

Modern cancer DNA sequencing tests generate vast amounts of data, with most commercial panels easily generating over 1 000 000 base pairs of data. These results must be filtered, interpreted, and presented to clinicians who will undertake medical decisions, frequently in the setting of multiple possible courses of action. Cancer Genome Interpreter is a software tool that streamlines and automates the process of identifying and annotating variants. The tool accepts several data formats and provides a user-friendly output. Also described is a new knowledge base of 5,314 validated mutations (the Catalog of Validated Oncogenic Mutations). As a proof of concept, 72% of AACR Project GENIE tumors (~17k) have at least one biomarker of drug response in the system. This proportion is much higher than what has been reported in older studies of genomically-informed treatment decisions, suggesting that the match between mutation and drug continues to improve.