Cancer Informatics in 2017: A New Beginning and a Bright Future

Jeremy L. Warner[1], Debra A. 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 2017.

Methods: An extensive search using PubMed/Medline, Google Scholar, and manual review was conducted to identify the scientific contributions published in 2017 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 three best papers was conducted by the editorial board of the Yearbook.

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

Conclusion: Cancer informatics is a broad and vigorous subfield of biomedical informatics. Strides in knowledge management, crowdsourcing, and visualization are especially notable in 2017.

Keywords

Neoplasms, informatics, health information technology, genomics, data visualization

Yearb Med Inform 2018:223-6
http://dx.doi.org/10.1055/s-0038-1667086

Introduction

The information being generated in the cancer clinic and the basic science lab is beginning to elucidate fundamental understandings of cancer beyond the traditional anatomically-driven approach. The nascent field of cancer informatics intends to take full advantage of the many data streams 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; and 3) manipulating the data, most commonly through visualization, to yield new insights. In this inaugural year of the Cancer Informatics section, we introduce the readers to a very broad and deep field which will continue to rapidly develop. As pointed out about cancer informatics research by Mathé, *et al.*, [1] in the survey paper of the Cancer Informatics section of this 2018 IMIA Yearbook, "*the cancer informatics revolution has been the beneficiary of a data explosion*". Big data being generated through genomics, metabolomics, and proteomics is the clearest avenue towards precision oncology.

In 2018, 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 2018 to identify peer-reviewed journal articles published in 2017, in the English language, related to cancer informatics research. In addition to the search through electronic databases, manual searches of key themes were performed in well-known informatics journals (e.g., Journal of the American Medical Informatics Association, Applied Clinical Informatics, Bioinformatics, Journal of Biomedical Informatics, etc.). Additionally, the contents of the journals JCO Clinical Cancer Informatics, JCO Precision Oncology, and Cancer Informatics were searched, as well as the contents of the 2017 special issue of Cancer Research: *Focus on Computer Resources*. For relevant articles that were PubMed indexed, we also searched for additional relevant articles using PubMed's "Similar articles" service. Finally, we also hand-searched the proceedings of MedInfo 2017, the 2017 AMIA Annual Symposium, and the 2017 AMIA Joint Summits.

One of the two section editors performed the searches. A PubMed search for the MeSH terms "Neoplasm" and "Informatics" yielded a total of 3,158 references. The Google Scholar search for "cancer informatics" returned 29,800 results. Given these vast results, we focused on identifying articles with translational or clinical applications, as opposed to 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 [2], the 15 candidate best papers were evaluated by the two section 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 2018 in the section 'Cancer Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Cancer Informatics
▪ Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, Chang MT, Chandarlapaty S, Traina TA, Paik PK, Ho AL, Hantash FM, Grupe A, Baxi SS, Callahan MK, Snyder A, Chi P, Danila D, Gounder M, Harding JJ, Hellmann MD, Iyer G, Janjigian Y, Kaley T, Levine DA, Lowery M, Omuro A, Postow MA, Rathkopf D, Shoushtari AN, Shukla N, Voss M, Paraiso E, Zehir A, Berger MF, Taylor BS, Saltz LB, Riely GJ, Ladanyi M, Hyman DM, Baselga J, Sabbatini P, Solit DB, Schultz N. OncoKB: a precision oncology knowledge base. JCO Precis Oncol 2017 Jul;2017.
▪ Newton Y, Novak AM, Swatloski T, McColl DC, Chopra S, Graim K, Weinstein AS, Baertsch R, Salama SR, Ellrott K, Chopra M, Goldstein TC, Haussler D, Morozova O, Stuart JM. TumorMap: exploring the molecular similarities of cancer samples in an interactive portal. Cancer Res 2017 Nov 1;77(21):e111-e114.
▪ Seyednasrollah F, Koestler DC, Wang T, Piccolo SR, Vega R, Greiner R, Fuchs C, Gofer E, Kumar L, Wolfinger RD, Winner KK, Bare C, Neto EC, Yu T, Shen L, Abdallah K, Norman T, Stolovitzky G, Soule HR, Sweeney CJ, Ryan CJ, Scher HI, Sartor O, Elo LL, Zhou FL, Guinney J, Costello JC, and Prostate Cancer DREAM Challenge Community. A DREAM Challenge to build prediction models for short-term discontinuation of docetaxel in metastatic castration-resistant prostate cancer. JCO Clin Cancer Inform 2017;1;1-15.

Conclusions and Outlook

The three selected best papers are representative of three distinct subdomains of cancer informatics: knowledge management, visualization, and crowdsourcing.

Chakravarty, *et al.,* [3] described a large public web resource, OncoKB (http://oncokb.org/) with a goal of providing evidence-based information about primarily somatic genomic variants for clinicians and researchers. The content is curated and stored in an internal data model which is exposed via a public application programming interface (API) on demand. As of this writing, OncoKB contains information on 477 genes, 3,855 variants, 60 tumor types, and 86 drugs. Twenty-five genes are linked to FDA-approved or standard of care treatment evidence, and 39 are linked to more limited clinical or biological (non-human) evidence. The practice of oncology is increasingly informed by biologic factors beyond the traditional biomarkers, in particular those obtained through clinical genomic sequencing; OncoKB is a cornerstone in the emergent ecosystem of cancer genomics knowledge management.

Newton, et al., [4] presented TumorMap (https://tumormap.ucsc.edu/), an interactive portal for the exploration of molecular similarities across cancer samples. TumorMap uses Google's Map technology to arrange samples in a hexagonal grid based on their similarity, after applying user-selected dimensionality reduction techniques. The resulting maps can be colored by various attributes such as clinical, molecular, phenotype, and outcome data and metadata. This novel and practical methodology allows not only the assessment of similarity and dissimilarity but also supports experimental research purposes. Outside of the clinical setting, high-dimensional -omics data are being routinely generated on large numbers of cancer samples and exploring these data in an intuitive and meaningful way is an unmet need that the authors have addressed.

Seyednasrollah, *et al.,* [5] described the process of the Prostate Cancer DREAM Challenge, the first crowd-sourced competition in metastatic prostate cancer and possibly in any cancer, which was launched in 2015; primary results are described elsewhere [6]. DREAM challenges are crowd-sourced competitions meant to accelerate progress in various biomedical informatics problems; importantly they are built on the FAIR data principles: Findable, Accessible, Interoperable and Reusable. This challenge focused on discontinuation of chemotherapy treatment for the reason of toxicity, which is as common as discontinuation for lack of efficacy. Predicting which patients are likely to experience early treatment failure is a critical issue in the oncology domain. The challenge attracted 34 independent teams from around the world and has led to post-challenge community collaborations. While the best results were only slightly better than reference, integrated time-dependent area under the curve (iAUC) of 0.791 versus 0.743, the successful operation of the challenge, which included data from four independent phase III randomized controlled trials, suggests a paradigm for crowd-sourced tasks in the oncology domain.

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

Huang, *et al.,* [7] and Kurnit, *et al.,* [8] described knowledge management approaches that bear some similarities with the OncoKB effort. Similar to the previously described Clinical Interpretation of Variants in Cancer (CIViC) [9] effort and MyCancerGenome [10], these knowledge bases help to build a rich ecosystem.

The American Association for Cancer Research (AACR) Project GENIE Consortium [11] described a substantial multi-institutional effort to aggregate next generation sequencing results obtained in routine clinical care with clinical annotations. The consortium recently expanded from eight to 19 institutions and the freely available data will become an increasingly valuable commodity for future research efforts.

Three of the candidate best papers [12–14] developed information extraction systems specific to the oncology domain. Bui, *et al.,* [12] focused on the extraction of oral chemotherapy exposure and Gao, *et al.,* [13] focused on pathology reports. In the tasks of primary site identification and histologic grade classification, hierarchical attention networks, a form of deep learning, achieve much better results than naïve Bayes, logistic regression, support vector machine, random forest, and other traditional machine learning models. Savova, *et al.,* [14] described DeepPhe, a

multipronged natural language processing approach to the extraction of cancer diagnosis and treatments across the longitudinal electronic medical record.

Hughes, *et al.,* [15] and Li, *et al.,* [16] recommended standards and guidelines for the interpretation and reporting of sequence variants in cancer and the health information technology needs of oncologists to facilitate the adoption of genomic medicine. These complementary guidelines by the American Society of Clinical Oncology, the Association for Molecular Pathology, and the College of American Pathologists provide a blueprint for the seamless integration of laboratory and clinical services for precision oncology.

Three of the candidate best papers [17–19] brought the patient and caregiver voices into cancer informatics, through the integration of patient-reported outcomes into routine cancer care, analysis of patient forum messages for signs of problems with medication adherence, and analysis of comment topics on a large cancer center's social media page.

Finally, Gandy, *et al.,* [20] described a software application for mining and presenting relevant cancer clinical trials per cancer mutation. Clinical trial matching in the era of precision oncology presents significant informatics challenges which are likely to be the focus of future editions of this chapter.

Acknowledgement

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

References

1. Mathé E, Hays JL, Stover DG, Chen JL. The omics revolution continues: the maturation of high-throughput biological data sources. Yearb Med Inform 2018:211-22.
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.
3. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol 2017 Jul;2017.
4. Newton Y, Novak AM, Swatloski T, McColl DC, Chopra S, Graim K, et al. TumorMap: exploring the molecular similarities of cancer samples in an interactive portal. Cancer Res 2017 Nov 1;77(21):e111–4.
5. Seyednasrollah F, Koestler DC, Wang T, Piccolo SR, Vega R, Greiner R, et al. A DREAM Challenge to build prediction models for short-term discontinuation of docetaxel in metastatic castration-resistant prostate cancer. JCO Clin Cancer Inform 2017 Aug 4;(1):1–15.
6. Guinney J, Wang T, Laajala TD, Winner KK, Bare JC, Neto EC, et al. Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data. Lancet Oncol 2017;18(1):132–42.
7. Huang L, Fernandes H, Zia H, Tavassoli P, Rennert H, Pisapia D, et al. The cancer Precision Medicine Knowledge Base for structured clinical-grade mutations and interpretations. J Am Med Inform Assoc 2017 May;24(3):513–9.
8. Kurnit KC, Bailey AM, Zeng J, Johnson AM, Shufean MA, Brusco L, et al. "Personalized Cancer Therapy": a publicly available precision oncology resource. Cancer Re. 2017 01;77(21):e123–6.
9. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet 2017 Jan 31;49(2):170–4.
10. Micheel CM, Lovly CM, Levy MA. My Cancer Genome. Cancer Genet 2014 Jun 1;207(6):289.
11. AACR Project GENIE Consortium. AACR Project GENIE: powering precision medicine through an international consortium. Cancer Discov 2017 Aug;7(8):818–31.
12. Bui N, Henry S, Wood D, Wakelee HA, Neal JW. Chart review versus an automated bioinformatic approach to assess real-world crizotinib effectiveness in anaplastic lymphoma kinase–positive non–small-cell lung cancer. JCO Clin Cancer Inform 2017 Mar 13;(1):1–6.
13. Gao S, Young MT, Qiu JX, Yoon H-J, Christian JB, Fearn PA, et al. Hierarchical attention networks for information extraction from cancer pathology reports. J Am Med Inform Assoc 2017 Nov 16;
14. Savova GK, Tseytlin E, Finan S, Castine M, Miller T, Medvedeva O, et al. DeepPhe: a natural language processing system for extracting cancer phenotypes from clinical records. Cancer Res 2017 Nov 1;77(21):e115–8.
15. Hughes KS, Ambinder EP, Hess GP, Yu PP, Bernstam EV, Routbort MJ, et al. Identifying health information technology needs of oncologists to facilitate the adoption of genomic medicine: recommendations from the 2016 American Society of Clinical Oncology Omics and Precision Oncology workshop. J Clin Oncol 2017 Sep 20;35(27):3153–9.
16. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer. J Mol Diagn 2017 Jan 1;19(1):4–23.
17. Tang C, Zhou L, Plasek J, Rozenblum R, Bates D. Comment topic evolution on a cancer institution's Facebook page. Appl Clin Inform 2017 Aug 23;8(3):854–65.
18. Wysham NG, Wolf SP, Samsa G, Abernethy AP, LeBlanc TW. Integration of electronic patient-reported outcomes into routine cancer care: an analysis of factors affecting data completeness. JCO Clin Cancer Inform 2017 Feb 22;(1):1–10.
19. Yin Z, Malin B, Warner J, Hsueh PY, Chen CH. The power of the patient voice: learning indicators of treatment adherence from an online breast cancer forum. Proceedings of the Eleventh International AAAI Conference on Web and Social Media; 2017. p. 337–46.
20. Gandy LM, Gumm J, Blackford AL, Fertig EJ, Diaz LA. A software application for mining and presenting relevant cancer clinical trials per cancer mutation. Cancer Inform 2017;16:1176935117711940.

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

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2018, Section Cancer Informatics

Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, Chang MT, Chandarlapaty S, Traina TA, Paik PK, Ho AL, Hantash FM, Grupe A, Baxi SS, Callahan MK, Snyder A, Chi P, Danila D, Gounder M, Harding JJ, Hellmann MD, Iyer G, Janjigian Y, Kaley T, Levine DA, Lowery M, Omuro A, Postow MA, Rathkopf D, Shoushtari AN, Shukla N, Voss M, Paraiso E, Zehir A, Berger MF, Taylor BS, Saltz LB, Riely GJ, Ladanyi M, Hyman DM, Baselga J, Sabbatini P, Solit DB, Schultz N

OncoKB: a precision oncology knowledge base

JCO Precis Oncol 2017 Jul;2017

The practice of oncology is increasingly informed by biologic factors beyond the traditional biomarkers, in particular those obtained through clinical genomic sequencing. Variants have prognostic and predictive implications and comprise a large and growing knowledge space. Chakravarty et al., have built a large public web resource, OncoKB (http://oncokb.org/) with a goal of providing evidence-based information for clinicians and researchers. The content is curated and stored in an internal data model which is exposed via a public API on demand. As of this writing, OncoKB contains information on 477 genes, 3,855 variants, 60 tumor types, and 86 drugs. Twenty-five genes are linked to FDA-approved or standard of care treatment evidence, and 39 are linked to more limited clinical or biological (non-human) evidence. OncoKB is a cornerstone in the emergent ecosystem of cancer genomics content management, and is developed and maintained by the Knowledge Systems group in the Marie Josée and Henry R. Kravis Center for Molecular Oncology at the Memorial Sloan Kettering Cancer Center (MSK), in partnership with Quest Diagnostics.

Newton Y, Novak AM, Swatloski T, McColl DC, Chopra S, Graim K, Weinstein AS, Baertsch R, Salama SR, Ellrott K, Chopra M, Goldstein TC, Haussler D, Morozova O, Stuart JM

TumorMap: exploring the molecular similarities of cancer samples in an interactive portal

Cancer Res 2017 Nov 1;77(21):e111-e114

Outside of the clinical setting, high-dimensional -omics data are being routinely generated on large numbers of cancer samples. This includes genomic, transcriptomic, proteomic, and epigenomic profiles that bring data to terabyte and petabyte scales. Exploring these data in an intuitive and meaningful way is an unmet need; despite the rapidly falling costs of technology, analysis remains a major bottleneck. In this Focus on Computer Resources article, the authors describe TumorMap (https://tumormap.ucsc.edu/), an interactive portal for the exploration of molecular similarities across cancer samples. TumorMap uses Google's Map technology to arrange samples in a hexagonal grid based on their similarity, after applying user-selected dimensionality reduction techniques. The resulting maps can be colored by various attributes such as clinical, molecular, phenotype, and outcome data and metadata. This novel and practical methodology allows not only the assessment of similarity and dissimilarity but also supports experimental research purposes.

Seyednasrollah F, Koestler DC, Wang T, Piccolo SR, Vega R, Greiner R, Fuchs C, Gofer E, Kumar L, Wolfinger RD, Winner KK, Bare C, Neto EC, Yu T, Shen L, Abdallah K, Norman T, Stolovitzky G, Soule HR, Sweeney CJ, Ryan CJ, Scher HI, Sartor O, Elo LL, Zhou FL, Guinney J, Costello JC, and Prostate Cancer DREAM Challenge Community

A DREAM challenge to build prediction models for short-term discontinuation of docetaxel in metastatic castration-resistant prostate cancer

JCO Clin Cancer Inform 2017 Aug 4;(1):1-15

Discontinuation of chemotherapy treatment for the reason of toxicity is as common as discontinuation for lack of efficacy. This may have major consequences when treatments have a very narrow therapeutic index. Predicting which patients are likely to experience early treatment failure is a critical issue in the oncology domain. One approach to this intractable problem is to engage the larger community to solve towards a common problem. The DREAM challenges are crowd-sourced competitions meant to accelerate progress in the resolution of various biomedical informatics problems; importantly they are built on the FAIR data principles: Findable, Accessible, Interoperable and Reusable. The article by Seyednasrollah, et al., describes the process of the Prostate Cancer DREAM Challenge, the first crowd-sourced competition in metastatic prostate cancer and possibly in any cancer, which was launched in 2015; primary results are described elsewhere. The challenge attracted 34 independent teams from around the world, and has led to post-challenge community collaborations. While the best results are only slightly better than reference, integrated time-dependent area under the curve (iAUC) of 0.791 versus 0.743, the successful operation of the challenge, which included data from four independent phase III randomized controlled trials, suggests a paradigm for crowd-sourced tasks in the oncology domain.