Published online: 2020-08-21



© 2020 IMIA and Georg Thieme Verlag KG

# Cancer Informatics in 2019: Deep Learning Takes Center Stage

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

[1] Departments of Medicine and Biomedical Informatics, Vanderbilt University, Nashville, TN, USA
[2] Texas Oncology, Austin, TX, USA

## Summary

**Objective**: To summarize significant research contributions on cancer informatics published in 2019.

**Methods**: An extensive search using PubMed/Medline and manual review was conducted to identify the scientific contributions published in 2019 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 committee of the Yearbook.

**Results**: The two selected best papers demonstrate the clinical utility of deep learning in two important cancer domains: radiology and pathology.

**Conclusion**: Cancer informatics is a broad and vigorous subfield of biomedical informatics. Applications of new and emerging computational technologies are especially notable in 2019.

## Keywords

Neoplasms, informatics, health information technology, deep learning, ethics

Yearb Med Inform 2020:243-6
http://dx.doi.org/10.1055/s-0040-1701993

## Introduction

Cancer informatics 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 research on cancer treatments; and 3) manipulating data to yield new insights. In this third year of the Cancer Informatics section of the International Medical Informatics Association (IMIA) Yearbook, we continue to focus on translational and clinical cancer informatics, with a special emphasis on ethics in concordance with the 2020 Yearbook theme. As pointed out by Griffin, *et al.*, [1] in the survey paper of the Cancer Informatics section of this IMIA Yearbook, "*while there are numerous innovations in the field of cancer informatics to advance prevention and clinical care, considerable challenges remain related to data sharing and privacy, digital accessibility, and algorithm biases and interpretation.*" In order to overcome these challenges, technology solutions cannot be considered in a vacuum, even those with very high performance.

In 2020, 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

One electronic database was searched, PubMed/MEDLINE. The search was performed in January 2020 to identify peer-reviewed journal articles published in 2019, in the English language, and 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 ("2019/01/01"[PDAT]: "2019/12/31"[PDAT]) AND English[lang])

This search yielded 3,323 results; the titles of all were manually reviewed by one of the two section editors, and the abstracts of 270 of these were manually reviewed by the same editor in order to arrive at a candidate list of 86 papers. The search was problematic for two reasons: 1) there was low specificity due to the frequent MeSH tagging of robotic surgery techniques, radiation oncology treatment planning, bioinformatics analyses, and conventional retrospective epidemiologic studies; and 2) content known to be in the clinical cancer informatics domain was not captured with high sensitivity. Despite these challenges, the theme of deep learning applications clearly emerged, especially in the realms of radiomics and pathomics.

For those papers reporting on a classification or prediction task, we generally took the performance measures into account when selecting the final 15 candidates, most commonly the area under the receiver operating characteristic curve (AUC). Both section editors classified the 86 candidate 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). 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.

## Conclusions and Outlook

The two selected best papers are both deep learning approaches in two important subspecialties for the field of cancer: radiology and pathology. This direction was anticipated by a National Cancer Policy Forum (NCPF) of the National Academy of Medicine workshop on Improving Cancer Diagnosis and Care held in 2018 [3].

Ardila, *et al.,* [4] describe a deep learning algorithm that uses a patient's current and prior computed tomography (CT) volumes to predict the risk of lung cancer. The model achieves 94.4% AUC on 6,716 National Lung Cancer Screening Trial cases [5] and performs similarly on an independent clinical validation set of 1,139 cases. Furthermore, the algorithm outperformed six expert radiologists with absolute reductions of 11% in false positives and 5% in false negatives. Lung cancer is the number one cancer killer and is felt to be much more curable if detected early, making this a major public health issue. Despite this, rates of CT lung cancer screening are low [6]. This study suggests one way in which the barrier to these low rates can be breached.

Campanella, et al., [7] developed a multiple instance learning-based deep learning system that uses only the reported pathologic diagnoses as labels for training. They evaluated the system on a very large single-institutional dataset comprising 44,732 whole slide images from 15,187 patients. Performance was evaluated on a limited number of cancer types: prostate cancer, basal cell carcinoma, and breast cancer metastatic to axillary lymph nodes. For these cancer types and circumstances, AUC was above 0.98, setting a clear new bar for performance of systems of this type. According to the authors, implementation of such a system in the clinical setting would allow pathologists to exclude 65-75% of slides while retaining 100% sensitivity. This type of automated performance could usher in a new era of pathology automation. While this study is very impressive, a controversy around the senior author's role in a for-profit venture with exclusive access to the whole slide images [8] raises some ethical questions.

The other candidate best papers cover the gamut of cancer informatics. Continuing on the theme of the selected best papers, Huang, *et al.,* [9] and Wong, *et al.,* [10] applied artificial intelligence techniques to cancer. Huang and colleagues used a convolutional neural network to determine BI-RADS category for breast ultrasound images. This is a hot topic area and while it missed the date cutoff for consideration in this Yearbook, the system described by McKinney, *et al.,* [11] sets a new standard benchmark. Wong and colleagues tackled a more circumscribed problem – the prediction of early biochemical recurrence of prostate cancer.

Three of the candidate best papers [12-14] tackle the challenge of the lack of standardization in many domains of cancer informatics in slightly different ways. Banerjee and colleagues use natural language processing (NLP) to detect breast cancer recurrence, an important concept with no commonly used structured correlate. Warner and colleagues introduce a standard terminology of chemotherapy regimens and related concepts. Xu and colleagues develop and validate an algorithm to detect breast cancer recurrence based on non-specific billing codes.

Wu, *et al.,* [15] and Kocak, *et al.,* [16] use radiomics approaches to predict genomic alterations in tumor tissue. This is a very interesting cross-over between disciplines and may accelerate a merging of the fields of molecular pathology and radiology, as envisioned by the NCPF report mentioned above. It will be interesting to follow the development of this new field of radiogenomics over the upcoming years.

Bernard, *et al.,* [17] and Lin, *et al.,* [18] develop interactive dashboards to elucidate the complexity of cancer. Bernard and colleagues describe a visualization technique to digest patient histories and illustrates this with the use case of post-operative prostate cancer. Lin and colleagues describe a multifaceted platform used to support studies on more than 50,000 patients with nasopharyngeal cancer.

Zuley, *et al.,* [19] and Maguire, *et al.,* [20] apply informatics methods to cancer registries. Information in registries is painstakingly collected through manual abstraction, and outside of the legally mandated registries there are a multitude of efforts to collect focused data, *e.g.*, the ACR National Mammography Database [21]. These papers describe efforts to link registries and to take advantage of free text fields using NLP.

Unlike in prior years of this section, only one knowledge base was selected as a finalist. Lever, *et al.,* [22] describe CancerMine, a literature-based resource of drivers, oncogenes, and tumor suppressors in cancer. The resource is freely available and downloadable at http://bionlp.bcgsc.ca/cancermine.

**Table 1** Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2020 in the section 'Cancer Informatics''. The articles are listed in alphabetical order of the first author's surname.

| Section |
| --- |
| Cancer Informatics |
| ▪ Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, Tse D, Etemadi M, Ye W, Corrado G, Naidich DP, Shetty S. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019 May 20;25:954-61. |
| ▪ Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019 Jul 15;25:1301-9. |

Finally, Zhu, *et al.,* [23] use NLP to identify social isolation affecting patients with prostate cancer. Applying informatics to social determinants of health is an excellent example of a positive application of ethics in informatics.

**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. Griffin AC, Topaloglu U, Davis S, Chung AE. From patient engagement to precision oncology: leveraging informatics to advance cancer care. Yearb Med Inform 2020; xxx.xxx
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. National Academies of Sciences E. Improving Cancer Diagnosis and Care: Clinical Application of Computational Methods in Precision Oncology: Proceedings of a Workshop [Internet]. 2019 [cited 2020 May 23]. Available from: https://www.nap.edu/catalog/25404/improving-cancer-diagnosis-and-care-clinical-application-of-computational-methods
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 May 20;25:954-61.
5. The National Lung Screening Trial: Overview and Study Design1. Radiology 2011 Jan;258(1):243–53.
6. Okereke IC, Nishi S, Zhou J, Goodwin JS. Trends in lung cancer screening in the United States, 2016–2017. J Thorac Dis 2019 Mar;11(3):873–81.
7. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019 Jul 15;25:1301-1309.
8. Ornstein C, Thomas K. Sloan Kettering's Cozy Deal With Start-Up Ignites a New Uproar. The New York Times [Internet]. 2018 Sep 20 [cited 2020 May 23]; Available from: https://www.nytimes.com/2018/09/20/health/memorial-sloan-kettering-cancer-paige-ai.html
9. Huang Y, Han L, Dou H, Luo H, Yuan Z, Liu Q, et al. Two-stage CNNs for computerized BI-RADS categorization in breast ultrasound images. BioMed Eng OnLine 2019;18:8.
10. Wong NC, Lam C, Patterson L, Shayegan B. Use of machine learning to predict early biochemical recurrence after robot-assisted prostatectomy. BJU Int 2019;123(1):51-7.
11. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577:89-94.
12. Banerjee I, Bozkurt S, Caswell-Jin JL, Kurian AW, Rubin DL. Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. JCO Clin Cancer Inform 2019 Oct;3:1-12.
13. Warner JL, Dymshyts D, Reich CG, Gurley MJ, Hochheiser H, Moldwin ZH, et al. HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. J Biomed Inform 2019;96:103239.
14. Xu Y, Kong S, Cheung WY, Bouchard-Fortier A, Dort JC, Quan H, et al. Development and validation of case-finding algorithms for recurrence fo breast cancer using routinely collected administrative data. BMC Cancer 2019;19:210.
15. Wu S, Meng J, Yu Q, Li P, Fu S. Radiomics-based machine learning methods for isocitrate dehydrogenase genotype prediction of diffuse gliomas. J Can Res Clin Oncol 2019;145:543-50.
16. Kocak B, Durmaz ES, Ates E, Ulusan MB. Radiogenomics in Clear Cell Renal Cell Carcinoma: Machine Learning–Based High-Dimensional Quantitative CT Texture Analysis in Predicting PBRM1 Mutation Status. AJR Am J Roentgenol 2019 Mar;212(3):W55-W63.
17. Bernard J, Sessler D, Kohlhammer J, Ruddle RA. Using dashboard networks to visualize multiple patient histories: a design study on post-operative prostate cancer. IEEE Trans Vis Comput Graph 2019;25(3):1615-28.
18. Lin L, Liang W, Li C-F, Huang X-D, Lv J-W, Peng H, et al. Development and implementation of a dynamically updated big data intelligence platform from electronic health records for nasopharyngeal carcinoma research. Br J Radiol 2019;92:20190255.
19. Zuley ML, Nishikawa RM, Lee CS, Burnside E, Rosenberg R, Sickles EA, et al. Linkage of the ACR National Mammography Database to the Network of State Cancer Registries: Proof of Concept Evaluation by the ACR National Mammography Database Committee. J Am Coll Radiol 2019 Jan;16(1):8–14.
20. Maguire FB, Morris CR, Parikh-Patel A, Cress RD, Keegan THM, Li C-S, et al. A text-mining approach to obtain detailed treatment information from free-text fields in population-based cancer registries: A study of non-small cell lung cancer in California. PLoS ONE 2019;14(2):e0212454.
21. National Mammography Database [Internet]. [cited 2020 May 23]. Available from: https://www.acr.org/Practice-Management-Quality-Informatics/Registries/National-Mammography-Database
22. Lever J, Zhao EY, Grewal J, Jones MR, Jones SJM. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat Methods 2019 Jun;16:505-7.
23. Zhu VJ, Lenert LA, Bunnell BE, Obeid JS, Jefferson M, Hughes-Halbert CA. Automatically identifying social isolation from clinical narratives for patients with prostate cancer. BMC Med Inform Dec Making 2019;19:43.

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: Content Summaries of Best Papers Selected for the 2020 Edition of the IMIA Yearbook, Section Cancer Informatics

### Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, Tse D, Etemadi M, Ye W, Corrado G, Naidich DP, Shetty S

### End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography

### Nat Med 2019 May 20;25:954-61

A deep learning algorithm that uses a patient's current and prior computed tomography volumes was developed to predict the risk of lung cancer. The model achieves 94.4% area under the curve (AUC) on 6,716 National Lung Cancer Screening Trial cases and performs similarly on an independent clinical validation set of 1,139 cases. Furthermore, the algorithm outperformed six expert radiologists with absolute reductions of 11% in false positives and 5% in false negatives. Lung cancer is the number one cancer killer and is felt to be much more curable if detected early, making this a major public health issue. Despite this, rates of CT lung cancer screening are low. This study suggests one way in which the barrier to these low rates can be breached.

### Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ

### Clinical-grade computational pathology using weakly supervised deep learning on whole slide images

### Nat Med 2019 Jul 15;25:1301-9

The authors developed a multiple instance learning-based deep learning system that uses only the reported pathologic diagnoses as labels for training. They evaluated the system on a very large single-institutional dataset comprising 44,732 whole slide images from 15,187 patients. Performance was evaluated on a limited number of cancer types: prostate cancer, basal cell carcinoma, and breast cancer metastatic to axillary lymph nodes. For these cancer types and circumstances, AUC was above 0.98, setting a clear new bar for performance of systems of this type. According to the authors, implementation of such a system in the clinical setting would allow pathologists to exclude 65-75% of slides while retaining 100% sensitivity. This type of automated performance could usher in a new era of pathology automation.