Journal of Radiation Oncology Informatics

A Step Forward in Cancer Informatics—It Is Mandatory to Make Guidelines Machine Readable

Nikola Cihoric[1*], Ivan Igrutinovic[2*], Alexandros Tsikkinis[1], Eugenia Vlaskou Badra[1], Paul-Henry Mackeprang[1]

1 Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
2 University of Kragujevac Faculty of Science, Kragujevac Serbia

**NC and II are equally contributing authors*

Correspondence: Nikola Cihoric MD, Inselspital, Bern University Hospital, Bern 3010, Switzerland
nikola.cihoric@gmail.com

Abstract Clinical guidelines are general recommendations for practicing clinicians regarding prevention, diagnosis and treatment of a given disease. One of the most comprehensive and used guidelines are developed and regularly updated by the National Comprehensive Cancer Network (NCCN). Guidelines are readily available for download in portable document format (PDF). A machine-readable representation of NCCN guidelines is currently not available. In this writing, we argue on the necessity that clinical guidelines should be published in a machine-readable format. After review of the available literature, we describe the most important achievements in the field. Publication of guidelines in a machine-readable form may also be beneficial for other scientific and technical disciplines.

Keywords: Clinical guideline | cancer treatment | machine readable documents | markup language | oncology informatics

Introduction

Clinical guidelines are general recommendations for practicing clinicians regarding prevention, diagnosis, and treatment of diseases. Ideally the recommendations are based on current high-level evidence. Methodological and practical directions for the development of guidelines are described for example in a publication from 2011 of the Institute of Medicine entitled “Clinical Practice Guidelines We Can Trust” [1].

The National Comprehensive Cancer Network (NCCN) Clinical Practice Guidelines

In oncology, one of the most comprehensive and used guidelines are those developed by the National Comprehensive Cancer Network (NCCN). NCCN is a non-profit alliance of 27 leading cancer centers based in the United States. The guidelines contain sequential management decisions and interventions and are applicable in the majority of clinical situations, covering 97% of cancer types affecting patients in the United States. They are continuously updated and revised, to stay current with the latest developments and evidence. They are an indispensable tool assisting physicians in decision making in cancer care.

After registration and acceptance of the terms and conditions, the NCCN guidelines are freely accessible for all interested parties in several forms. Namely, users have access to NCCN Guidelines with NCCN Evidence Blocks™, NCCN International Translations/Adaptations, NCCN Educational Events and Programs and the core NCCN Clinical Practice Guidelines. The newest form is the NCCN Framework for Resource Stratification of NCCN Guidelines (NCCN Framework™), subclassified as basic, core and enhanced. Altogether, the guidelines are available to download in the commonly used portable document format (PDF). The PDF format is designed with the intention to represent textual and graphical data across multiple platforms in the way they are supposed to be viewed or printed. PDF is not intended for structured or semi-structured data exchange, data extraction or mining. Content extraction from a PDF file in a structured way is not impossible, but it is also not a straightforward task. The complete PDF documentation file released from Adobe extends over 900 pages [2]. Although some support for data mining exists, the complexity of the format is prohibitive, if not impossible to manage, for the majority of scientists and scientific software developers.

Data and information contained within the NCCN guidelines are invaluable. They are structured, interconnected and represented with workflow graphics and text. They are properly referenced to the source literature and

enriched with a metadata system in the form of evidence levels. As such, they represent a unique resource and it is imperative to make an additional effort and transform this amalgamated knowledge into a form suitable for informatics approaches. There are multiple reasons for this, but the most important ones can be summarized as follows.

Machine readable medical guidelines – a necessity

Firstly, we are confronted with an explosion of information. As the average life expectancy increases the total number of patients is also rising [3]. The introduction of new services will provide more opportunities to collect and analyze data [4, 5]. For every patient we will have more and more prognostic and predictive data which our decisions should be based on, especially factors based on "–omics" data [6]. Tumorboards based on molecular profiling of cancers are already being introduced into our practice and will play an ever more important role as we go forward [7-9]. Scientific output is growing with significantly [10], followed by a rising number of therapeutic options [11]. This information overload, as well as the inadequate use of technology modern technology does not come without its price [12, 13]. It is evident that it will be ever harder to keep up with developments. Modern tools for assistance and facilitation of cancer care are mandatory [14, 15]. As an example, development and maintenance of decision support tools for oncology would be easier. The implementation of decision tools based on structured NCCN guidelines is almost self-explanatory. Significant parts of the NCCN guidelines are expressed with branching logic where conditional statements are given in form of information and sugesstions or action interconnected with if-then-else logic.

Secondly, the information growth is accompanied with an even faster expansion of scientific publications [10]. The rise in quantity is usually not accompanied with a rise in quality [16]. To search for relevant and high-quality information we need to spend extensive time and energy. Text mining of cancer related data is rather underdeveloped, and efforts in this direction are more incidental than systematic [17]. We could use a well-structured database of evidence-based papers and recommendations for evaluating existing and new literature and for machine learning processes. If we want to develop machine learning techniques for evidence classification based on natural language processing we will need training material, and it is hard to imagine a better one than the structured NCCN guideline.

Thirdly, treatment quality and conformance to best practice was and is still an issue [18-20]. Quality control and conformance to standards would be facilitated through early feedback on structure and content. It is possible to imagine that the National Guideline Clearinghouse would embrace such undertaking.

However, the idea of a structured approach to clinical guidelines is not new [21]. Shahar et al. described a text-based language for representation and annotation of clinical guidelines (CG) [22]. A few years later, Shahar et al published an interesting research paper on the efforts to convert guidelines written as a free text to annotated ontology enriched digital electronic guidelines [23]. This approach should be embraced in terms of research in natural language processing. But for practical purposes, we should think more straightforwardly and provide the desired results directly from the source. As a matter of fact, more effort is given to creating structure from unstructured text with complicated and complex approaches than to establishing a well-defined structure to begin with. The same group has developed the Digital Electronic Guidelines Library (DeGeL), a web-based guideline repository and a suite of tools, to support the use of automated guidelines for medical care, research, and quality assessment [24]. Several authors have evaluated or proposed different approaches for machine readable implementation or development of guidelines. Johnson et al. describe PRODIGY, a guideline-based decision support system aimed at the support of general practitioners [25]. Tu and Musen have also described a task oriented approach to guideline modeling developed within the EON project [26]. Guidelines seek to change behavior by making statements involving one or all tasks: setting constraints, setting goals, making decisions, sequencing and synchronization of actions, interpreting the data [26]. Peleg et al. developed the Guideline Interchange Format Language (GLIF) which has evolved through several versions [27]. GLIF consists of three levels, namely Conceptual, Computable and Implementable Level. The conceptual level of GLIF was described with the Unified Markup Language. Computable and implementable layers allow some of the higher programming concepts such as macros. Although interesting, none of the described models and proposals have been broadly implemented in practice. The Arden syntax for medical logic modules is an industry recognized standard for expressing medical knowledge. However, broader utilization in terms of guideline representation may be prohibitive due to technological limitations and complexity [28-31].

The data in NCCN guidelines should be available in standards intended for data exchange that are both readable for machines and humans. The format should be simple and understandable for a broad audience. It does not have to fulfill any other requirements but solely focus on representing information contained in clinical guidelines in a structured way. An excellent candidate is the extensible markup language (XML) developed and managed by World Wide Web Consortium (W3C). In summary, XML is a markup language intended for textual data processing in a semi-structured way. It is easily understandable and widely implemented in practice. It is used not only for textual data but for any kind of data requiring a structured approach for processing and exchange. The XML structure is described through XML Schemas (XSD). The clinicaltrials.gov portal and the Clinical Data Interchange Standard Consortium

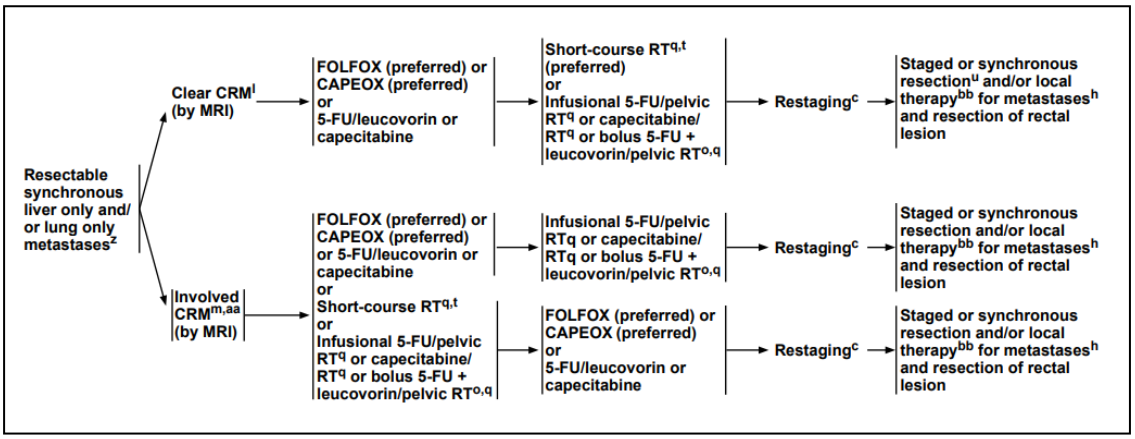

Figure 1. Example of NCCN guideline workflow diagram for metastatic rectal cancer.

(CDISC) use XML as one of the formats for data exchange and their schemas ae free and publicly available. XML facilitates the development of tools such as syntax checkers and editors that can help increase the correctness of the content, and this in turn will foster the development and production. XML has also invaluable extensions in the form of XPath and XQuery which are Turing complete (A computational system that can compute every Turing-computable function). However, one can use other machine readable formats such as javascript object notation (JSON), hypertext markup language (html) or develop a new standard. However, this would be connected with extensive additional work for tool development, syntax checkers and parsers.

Furthermore, NCCN guidelines contain specific workflow rules expressed with IF – THEN control flow (Figure 1).

The limitations of our proposal should be acknowledged. We have concentrated our research and discussion only on the NCCN guidelines, based on the fact that the consortium was established with the main purpose of guideline development and maintenance. We cannot exclude the possibility that other parties have already developed and established a machine-readable guideline system. However, a brief survey on the major oncological societies which publish clinical practice guidelines on a regular basis did not yield any documents available in the proposed format (ASCO, EORTC, ASTRO, ESTRO, AGO). We did not give any specific recommendation for further development in terms of XML schema content, but this will certainly be the topic of future publications.

Conclusion

Guidelines in general should be available in a machine readable form. The format should be utilized in scientific efforts and implemented into clinical routine. On a technical level, it is possible to imagine the integration of such resources into decision support systems or quality assurance audits. However, the intention of this paper is not to discuss lower technical aspects of implementation, but to raise awareness and motivate the responsible consortia and the scientific community. The same resources, made machine readable, can improve the fight against cancer and would will certainly be welcomed by the scientific community.

References

1. In: Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E, editors. Clinical Practice Guidelines We Can Trust. Washington (DC)2011.

2. Incorporated AS. Document Management - Portable Document Format - Part 1 [Internet]. Internet: Adobe Systems Incorporated; 2008 [cited 2018 08.02.2018]. Available from: https://www.adobe.com/content/dam/acom/en/devnet/pdf/pdfs/PDF32000_2008.pdf.

3. Smith BD, Smith GL, Hurria A, Hortobagyi GN, Buchholz TA. Future of cancer incidence in the United States: burdens upon an aging, changing nation. J Clin Oncol. 2009;27(17):2758-65. doi: 10.1200/JCO.2008.20.8983. PubMed PMID: 19403886.

4. Schilsky RL, Michels DL, Kearbey AH, Yu PP, Hudis CA. Building a rapid learning health care system for oncology: the regulatory framework of CancerLinQ. J Clin Oncol. 2014;32(22):2373-9. doi: 10.1200/JCO.2014.56.2124. PubMed PMID: 24912897.

5. Shah A, Stewart AK, Kolacevski A, Michels D, Miller R. Building a Rapid Learning Health Care System for Oncology: Why CancerLinQ Collects Identifiable Health Information to Achieve Its Vision. J Clin Oncol. 2016;34(7):756-63. doi: 10.1200/JCO.2015.65.0598. PubMed PMID: 26755519.

6. Waldron D. Cancer genomics: A multi-layer omics approach to cancer. Nature reviews Genetics. 2016;17(8):436-7. doi: 10.1038/nrg.2016.95. PubMed PMID: 27418154.

7. van der Velden DL, van Herpen CML, van Laarhoven HWM, Smit EF, Groen HJM, Willems SM, et al. Molecular Tumor Boards: current practice and future needs. Ann Oncol. 2017;28(12):3070-5. doi: 10.1093/annonc/mdx528. PubMed PMID: 29045504.

8. Parker BA, Schwaederle M, Scur MD, Boles SG, Helsten T, Subramanian R, et al. Breast Cancer Experience of the Molecular Tumor Board at the University of California, San Diego Moores Cancer Center. J Oncol Pract. 2015;11(6):442-9. doi: 10.1200/JOP.2015.004127. PubMed PMID: 26243651.

9. Bardia A, Iafrate JA, Sundaresan T, Younger J, Nardi V. Metastatic Breast Cancer With ESR1 Mutation: Clinical Management Considerations From the Molecular and Precision Medicine (MAP) Tumor Board at Massachusetts General Hospital. Oncologist. 2016;21(9):1035-40. doi: 10.1634/theoncologist.2016-0240. PubMed PMID: 27551012; PubMed Central PMCID: PMCPMC5016066.

10. Larsen PO, von Ins M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics. 2010;84(3):575-603. doi: 10.1007/s11192-010-0202-z. PubMed PMID: 20700371; PubMed Central PMCID: PMCPMC2909426.

11. Mullard A. 2017 FDA drug approvals. Nat Rev Drug Discov. 2018;17(2):150. doi: 10.1038/nrd.2018.18. PubMed PMID: 29386602.

12. Wright AA, Katz IT. Beyond Burnout - Redesigning Care to Restore Meaning and Sanity for Physicians. N Engl J Med. 2018;378(4):309-11. doi: 10.1056/NEJMp1716845. PubMed PMID: 29365301.

13. Dzau VJ, Kirch DG, Nasca TJ. To Care Is Human - Collectively Confronting the Clinician-Burnout Crisis. N Engl J Med. 2018;378(4):312-4. doi: 10.1056/NEJMp1715127. PubMed PMID: 29365296.

14. Karsh BT, Weinger MB, Abbott PA, Wears RL. Health information technology: fallacies and sober realities. J Am Med Inform Assoc. 2010;17(6):617-23. doi: 10.1136/jamia.2010.005637. PubMed PMID: 20962121; PubMed Central PMCID: PMCPMC3000760.

15. Hesse BW, Ahern D, Beckjord E. Oncology informatics: Using health information technology to improve processes and outcomes in cancer: Academic Press; 2016.

16. Smith R. The trouble with medical journals. Journal of the Royal Society of Medicine. 2006;99(3):115-9.

17. Spasic I, Livsey J, Keane JA, Nenadic G. Text mining of cancer-related information: review of current status and future directions. Int J Med Inform. 2014;83(9):605-23. doi: 10.1016/j.ijmedinf.2014.06.009. PubMed PMID: 25008281.

18. Winn RJ. Oncology practice guidelines: do they work? Journal of the National Comprehensive Cancer Network. 2004;2(4):276-82.

19. Somerfield MR, Einhaus K, Hagerty KL, Brouwers MC, Seidenfeld J, Lyman GH. American Society of Clinical Oncology clinical practice guidelines: opportunities and challenges. Journal of Clinical Oncology. 2008;26(24):4022-6.

20. Kung J, Miller RR, Mackowiak PA. Failure of clinical practice guidelines to meet institute of medicine standards: two more decades of little, if any, progress. Archives of internal medicine. 2012;172(21):1628-33.

21. Balas EA, Puryear J, Mitchell JA, Barter B. How to structure clinical practice guidelines for continuous quality improvement? J Med Syst. 1994;18(5):289-97. PubMed PMID: 7861105.

22. Shahar Y, Miksch S, Johnson P. An intention-based language for representing clinical guidelines. Proc AMIA Annu Fall Symp. 1996:592-6. PubMed PMID: 8947735; PubMed Central PMCID: PMCPMC2233124.

23. Shahar Y, Shalom E, Mayaffit A, Young O, Galperin M, Martins S, et al., editors. A distributed, collaborative, structuring model for a clinical-guideline digital-library. AMIA Annual Symposium Proceedings; 2003: American Medical Informatics Association.

24. Shahar Y, Young O, Shalom E, Mayaffit A, Moskovitch R, Hessing A, et al. The Digital electronic Guideline Library (DeGeL): a hybrid framework for representation and use of clinical guidelines. Stud Health Technol Inform. 2004;101:147-51. PubMed PMID: 15537218.

25. Johnson PD, Tu S, Booth N, Sugden B, Purves IN. Using scenarios in chronic disease management guidelines for primary care. Proc AMIA Symp. 2000:389-93. PubMed PMID: 11079911; PubMed Central PMCID: PMCPMC2244127.

26. Tu SW, Musen MA. A flexible approach to guideline modeling. Proc AMIA Symp. 1999:420-4. PubMed PMID: 10566393; PubMed Central PMCID: PMCPMC2232509.

27. Peleg M, Boxwala AA, Ogunyemi O, Zeng Q, Tu S, Lacson R, et al. GLIF3: the evolution of a guideline representation format. Proc AMIA Symp. 2000:645-9. PubMed PMID: 11079963; PubMed Central PMCID: PMCPMC2243832.

28. Hripcsak G. Arden Syntax for Medical Logic Modules. MD Comput. 1991;8(2):76, 8. PubMed PMID: 2038238.

29. Kuhn RA, Reider RS. A C++ framework for developing Medical Logic Modules and an Arden Syntax compiler. Comput Biol Med. 1994;24(5):365-70. PubMed PMID: 7705067.

30. Jenders RA, Dasgupta B. Assessment of a knowledge-acquisition tool for writing Medical Logic Modules in the Arden Syntax. Proc AMIA Annu Fall Symp. 1996:567-71. PubMed PMID: 8947730; PubMed Central PMCID: PMCPMC2233222.

31. Choi J, Bakken S, Lussier YA, Mendonca EA. Improving the human readability of Arden Syntax medical logic modules using a concept-oriented terminology and object-oriented programmingexpressions. Comput Inform Nurs. 2006;24(4):220-5. PubMed PMID: 16849918; PubMed Central PMCID: PMCPMC2883181.