Curr. Treat. Options in Oncol. (2018) 19:62
DOI 10.1007/s11864-018-0585-2

Head and Neck Cancer (L Licitra, Section Editor)

Big Data in Head and Neck Cancer

Carlo Resteghini, MD[1,*]
Annalisa Trama, MD, PhD[2]
Elio Borgonovi, BA, PhD[3]
Hykel Hosni, PhD[4]
Giovanni Corrao, PhD[5]
Ester Orlandi, MD[6]
Giuseppina Calareso, MD[7]
Loris De Cecco, PhD[8]
Cesare Piazza, MD[9,10]
Luca Mainardi, PhD[11]
Lisa Licitra, MD[1,10]

Address
[*,1]Head and Neck Medical Oncology Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
Email: carlo.resteghini@istitutotumori.mi.it
[2]Evaluative Epidemiology Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[3]Department of Policy Analysis and Public Management, Research Center on Health and Social Care Management, CERGAS, SDA Bocconi School of Management, Bocconi University, Milan, Italy
[4]Department of Philosophy, University of Milan, Milan, Italy
[5]Department of Statistics and Quantitative Methods, Division of Biostatistics, Epidemiology and Public Health, University of Milano-Bicocca, Milan, Italy
[6]Radiotherapy 2 Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Via Giacomo Venezian, 1, 20133, Milan, MI, Italy
[7]Department of Radiology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[8]Integrated Biology Platform, Department of Applied Research and Technology Development, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[9]Department of Otorhinolaryngology, Maxillofacial, and Thyroid Surgery, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[10]University of Milan, Milan, Italy
[11]Department of Electronic, Information, and Bioengineering, Politecnico di Milano, Milan, Italy

This article is part of the Topical Collection on *Head and Neck Cancer*

Keywords Big data · Support vector machine · Machine learning · Head and neck cancer · Genomics · Radiomics · Surgery · Oncology · Forecasting · Evidence based medicine · Guidelines · Decision support system

Published online: 25 October 2018

Opinion statement

Head and neck cancers can be used as a paradigm for exploring "big data" applications in oncology. Computational strategies derived from big data science hold the promise of shedding new light on the molecular mechanisms driving head and neck cancer pathogenesis, identifying new prognostic and predictive factors, and discovering potential therapeutics against this highly complex disease. Big data strategies integrate robust data input, from radiomics, genomics, and clinical-epidemiological data to deeply describe head and neck cancer characteristics. Thus, big data may advance research generating new knowledge and improve head and neck cancer prognosis supporting clinical decision-making and development of treatment recommendations.

Introduction

Technological innovation combined with automation has generated a tremendous amount of available data, also called "big data" [1, 2]. However, what big data means for health care is not straightforward. A recent review [3] proposed to define big data exclusively by their volume, considering "big dataset" only if $\mathrm{Log}^{(n \times p)}$ is superior or equal to 7, where n is the number of individuals and p the number of variables. This was proposed considering that problems related to computational methods do not exist for $\mathrm{Log}^{(n \times p)}$ inferior to 7.

Alternative widely used notions to describe the big data are the three "Vs": volume, variety, and velocity [3] or that they are so large and complex that they are difficult to manage with traditional software and/or hardware, nor can they be easily managed with traditional or common data management tools and methods [4].

Therefore, the previous are management-centered definition that highlights the core of innovation brought by big data: a novel way to interrogate (large) amount of information.

In this article, we refer to the definition proposed during the workshop "Big data in health research: An EU action plan" [5•] according to which "Big data in health" encompasses high volume, high diversity biological, clinical, environmental, and lifestyle information collected from single individuals to large cohorts, in relation to their health and wellness status, at one or several time points. Thus, big data comes from a variety of sources, such as clinical trials; electronic health records; patient registries and databases; multidimensional data from genomic, epigenomic, transcriptomic, proteomic, metabolomic, and microbiomic measurements; and medical imaging. More recently, data are being integrated from social media, socioeconomic or behavioral indicators, occupational information, mobile applications, or environmental monitoring [6].

This article reviews the use of big data and its related technology to study head and neck cancers (HNCs). New cases in Europe are expected to affect approximately 151,000 persons in 2020 [7]. However, HNCs are rare when considered by anatomical subsites and histotypes (incidence rate < 6/100,000) [8].

To contribute to the discussion on the use of big data in health care and research, we focused on the following objectives:

- to understand to what extent big data are used in the field on head and neck oncology,
- to identify the purposes for which big data are currently used in the field of head and neck oncology,
- to understand the approach used (traditional vs. big data analytics) for analyzing big data.

Applications of big data technology in HNCs

Results of literature revision reflect the innovative nature of big data approach in this field, with several works published in recent years focusing on five main categories.

Genomics

Here, we focus our interest not simply on genomic analysis in HNCs, which is a well-established field of research, but on application of analytics technology to genomic data. In particular, the use of support vector machine (SVM) [9] is a trending approach in several cancers [10], and HNC research is proceeding in this line. This machine learning method has been employed to generate prognostic prediction model analyzing genomic data. Results have been reported for laryngeal [11, 12], nasopharyngeal [13], oral cavity [14, 15], and HPV-positive oropharyngeal carcinoma [16] and mixed subsites of HNCs treated with postoperative chemoradiation [17].

Radiomics

The term radiomics refers to the application of a large number of quantitative image features to obtain the comprehensive quantification of tumor phenotypes [18]. It represents the most explored field of analytics investigation in HNCs.

Computer tomography (CT) imaging has been the focus of the early works in the radiomics research [18]. In HNC field, the routine non-contrast-enhanced CT scan performed for radiotherapy planning constitutes the basis for radiomics development. Examples are especially aimed to provide a prognostic biomarker in HNCs [19, 20, 21•] and in specific subsites, such as the oropharynx [22, 23]. This particular subsite has been the focus of another investigation exploiting CT-based radiomics to detect HPV status [24].

Other future applications of radiomics could impact intraoperative procedures [25] and provide toxicities and treatment response prediction models [26].

The advances in this field are underlined by the availability of an already published review [27•] focused on MRI radiomics; we refer the reader to this paper for further insight on this topic. In this rapidly evolving field of investigations, MRI provides a high number of features suitable for diagnostic and prognostic purposes, especially with functional sequences such as *diffusion-weighted imaging* (DWI) and *apparent diffusion coefficients* (ADC). Examples are represented by an MRI-based tool to differentiate sinonasal malignancy from inverted papilloma [28] or the use of MRI radiomics analysis as prognostic indicator in nasopharyngeal carcinomas [29, 30].

We were able to identify only few papers exploring HNC's radiomics employing FDG-PET/CT scan, aiming to elaborate an algorithm for automatic diagnosis of nasopharyngeal carcinomas [31] and to predict outcome based on pre- and post-radio(chemo)therapy treatment in oropharyngeal carcinomas [32].

Optical imaging and enhancement technology

Clinical and optical examination is pivotal in patient's evaluation and the conduction of surgical procedures. Optical imaging technology has developed using different light sources or bioactive enhancers [33, 34]. Nevertheless, these advancements do not involve big data technologies. As much as with radiological data, images generated during a clinical examination are suitable for elaboration by algorithms able to enhance clinician's diagnostic performance. Few examples of this application are available in the literature, evaluating both preoperative [35–40] and intraoperative [41, 42] settings. Simulators for training and presurgical planning employing virtual reality are under development and could be useful in the near future for both clinical and educational purposes [43, 44].

Decision support tools able to distinguish precancerous—such as leukoplakias—from cancerous lesions are under evaluation [45, 46].

Radiation therapy

Radiotherapy is one of the most effective and employed treatments for HNCs. Its aim is to achieve a high probability of local tumor control at a low risk of normal tissue complications. Normal tissue complication probability (NTCP) is function of a complex of factors such as late radiation responses, tumor control, and toxicities [47]. Analytics and big data technologies have been employed to develop a decision-support system to model treatment outcome and NTCP. These methodologies take into account the growing complexity of radiation therapy, combining both predictive and prognostic data factors from clinical, imaging, molecular, and other sources to achieve the highest accuracy to predict tumor response and follow-up event rates [48].

Our literature review documents several efforts to define NTCP model for HNC radiotherapy employing big data analytics [49–54]. Other fields of interest seem to be related to adaptive radiotherapy [55, 56] and automated treatment planning [57, 58].

Miscellaneous and integrations

The intrinsic complexity of HNCs and its treatments implies that multiple factors drive patient's outcome. Such complex of heterogeneous, non-genetic factors has been referred to as exposome [59]. Emerging technologies provide detailed information on drugs, toxicants, pollutants, nutrients, and physical and psychological stressors on an omics scale. Examples of exposome investigation in HNCs are available [60], but still lack a big data analytic approach.

Big data analysis has been employed in order to predict oncologic prognosis, focusing on several inputs such as administrative records [61], hematological markers and clinical characteristics [62], treatment specifics, or combinations of several data [63, 64].

We report other fields of intriguing prospectives that may represent future breakthroughs of big data technology. Some has described the possibility to anticipate the potential activity of drugs in HNCs [65, 66], while others are evaluating the feasibility of automated tumor aggressiveness assessment based on pathological hematoxylin and eosin-stained slides [67]. All previously cited works present interesting prospectives with promising results. Anyway, they focus on a singular aspect of big data technology, limiting their potential. Therefore, we also report on what we consider to be the next step in this field of research. A multicentric study (*Big Data and Models for Personalized Head and Neck Cancer Decision Support, BD2Decide*) [68•] is ongoing in Europe to develop a decision support system for the prediction of HNC patient disease outcomes, in particular for advanced stages. This will be obtained by new multiscale signatures integrating radiomics, genomics, and clinical-epidemiological data.

Current status of big data research in HNCs

We developed a brief questionnaire about the use of big data in HNC research based on the following steps:

- review of the literature,
- development of a draft questionnaire,

– discussion of the draft questionnaire with different experts including statisticians, research methodologists, epidemiologists, economists, and experts in logic.

The questionnaire was sent to major international collaborative research groups working on HNCs and to pharmaceutical companies supporting HNC research for a total of about 100 key research players.

Sixty-one, out of 100, responded to the questionnaire (response rate = 60%). Responders' understanding of big data was in line with the one used in this article. Thus, the main characteristics of big data enlisted by responders included the following: big sample size, data variability, and the need of algorithmic analyses. Only half of the responders were involved in big data research on HNC objectives. The data sources mostly used in the "big data" studies were clinical reports (either paper or electronic), radiologic and genomic data, and, to a lesser extent, population-based cancer registries. These three sources were also ranked as the most relevant. Data quality issues were considered by most of the responders a relevant concern because the quality of health care data is highly variable and often incorrect; data are implausible; there is a redundancy of data entries and duplicates; and results coming from different laboratories, diagnostic techniques, and devices have to be carefully managed. Thus, most of responders (68%) had planned to validate the results either with prospective observational studies or with clinical trials.

Objectives of big data research in HNCs

Table 1 describes the aims of the studies responders are currently undertaking on HNC using big data.

The aims listed were heterogeneous but included both descriptive and analytical objectives (Fig. 1). Interestingly, regardless of different objectives, nearly all responders agreed that the main real-life applications for the ongoing big data research were the development of tools to:

– implement decision-making process in areas of uncertainty,
– define clinical guidelines,
– read complex phenomenon and generate new hypotheses.

Table 1. Aims of HNC big data studies according to survey participants

Answer choices	Responders
Identification of risk factors/profiles	60.0%
Support to clinical decision-making process	56.7%
Safety and effectiveness assessment of a given therapeutic strategy in a real-life setting	33.3%
Developing tools, methods and algorithms for generating real-world evidence from real-world (big)data	36.7%
Assessing health care pathways and evaluating the corresponding appropriateness, costs, etc.	13.3%
Measuring risk-benefit and/or cost-effectiveness profiles of health care pathways	10.0%
Other	20.0%

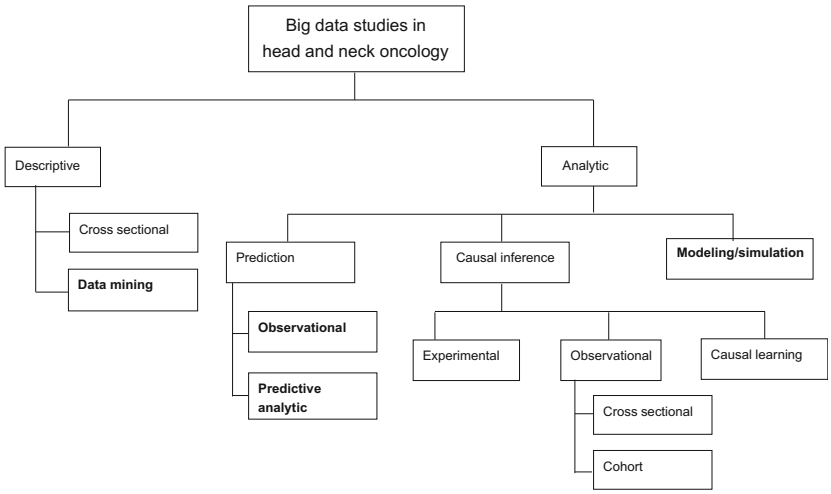

Fig. 1. Approach to big data analysis. This figure shows the type of big data studies currently ongoing on HNCs based on the taxonomy proposed by Sim [69••]. Based on survey results and literature review, the ongoing studies are mainly classifiable as observational studies analyzed with a traditional approach.

These applications represent areas of unmet needs for HNC treatment. Given the nature of rare and heterogeneous tumor subtypes affecting patients with multiple comorbidities, the possibility to deliver high-quality treatment is impaired by lack of evidences. Big data could provide insights in the complexity of HNCs, generating evidences for guidelines and personalized treatments driving clinicians through uncertainties.

Future directions: how to exploit big data technology

Our review showed that big data are not commonly used in HNC research. Studies based on big data are limited in numbers and based on traditional types of study design and analysis. However, some specific big data applications, namely radiomics and genomics, attract increasing interests.

Indeed, radiomics provides informative data for radiotherapic dose planning and toxicity prevention, while, until now, HNC genomic research did not provide targets suitable of therapy and the main classifier for HNCs appears to be HPV status rather than a specific genetic profile [70]. Big data technology could facilitate the comprehension of the complex interaction of tumor's and host's biology. An increased understanding at this level would enable to generate useful tools for actual precision medicine application.

Support to clinical decision

HNCs are composed of an heterogeneous group of complex diseases, often with atypical presentations. Multidisciplinary (MDT) treatments require intense and specialized supportive care and high expertise which is lacking in most hospitals, given HNC low incidence rate [71]. Thus, HNC challenges the current cancer care model—based on trained physicians that shape solutions for the single patient [72•]—because of the disease, patient, and treatment complexity and the shortage of expertise.

A new model for care delivery based on human-machine collaboration was proposed [72•]. In our opinion, this model can contribute to address the HNC complexity because it will create a system for decision support. This model is based on three components: software-based algorithms, physician innovation collaboratives (PICs), and clinician mix optimization. In our opinion, the first two components are those more relevant for HNCs.

Next-generation software-based algorithms are emerging, e.g., CancerLinQ and Watson. These algorithms enable rapid learning from patient data and could allow for continuous adjustments to treatment plans [72•]. In addition, emerging technologies will progressively handle not structured and diverse data, furthering the potential of algorithms application [73, 74]. However, softwares alone are not enough. Physicians have to coach their software-based algorithms which means that physicians have to commit to using the same software-based algorithms in their practices, to assess the performance of and adjust the algorithms [72•]. Any adjustments would be tested and refined on new cohorts of patients, creating a feedback loop that would be continuously iterated over time [72•]. This process is well-suited to cancer care and especially to HNCs since physicians dealing with such an oncological subspecialty already have a practice of MDT team meetings through their tumor boards.

We do not believe, however, that the proposed clinician mix optimization (i.e., appropriately supervised and specialty-trained nurse practitioners or physician assistants executing software-based algorithms at the point of care to deliver it in a better or more efficient way) would be appropriate for HNCs [72•]. This is due to the HNC characteristics which require the active involvement of experts from a particularly wide variety of fields. An MDT for HNCs should include at least a head and neck surgeon, a radiation oncologist, a pathologist, a radiologist, and a medical oncologist. For HNCs, the MDT has to take overall responsibility for assessment, treatment planning, and management of all patients throughout the course of their disease including decision about supportive care and rehabilitation. Thus, in our opinion, it seems unlikely to substitute the figure and role of contemporary MDT. On the opposite, it seems very relevant the idea to provide the MDTs with tools to support and implement their decisions, to ameliorate their performance and ultimately the patients' outcomes.

Finally, the use of software-based algorithms in the clinical practice must be paired with robust methods to enlighten their relevant aspects [75]. This is particularly relevant for rare pathologies such as HNCs in which spurious correlations with uncommon conditions may lead to inaccurate classifications [76]. More importantly, new training will be needed to provide physicians with the conceptual skills to interpret the outputs of algorithmic decision support systems in light of the known challenges [76, 77].

Use of big data for generating new knowledge

Knowledge is considered the ultimate result of the process that elaborates crude data with information and context. For example, a raw data as *p16 immunohistochemistry positivity* is somehow useless without proper context. Adding information like patient's clinical characteristics, biopsy site, and pathologic diagnosis (e.g., TxN1M0 squamous cell carcinoma from a level III neck lymph node) defines the context of the p16 positivity data and generates knowledge. Therefore,

deeper disease understanding (e.g., high probability of oropharyngeal primary site) and more accurate prediction (good overall prognosis) can be drawn.

Knowledge is the driving force behind reasonable action, and medicine is no exception. Data science promise is to allow for different views and insights of known problems though artificial intelligence (AI, i.e., machine learning, deep learning) generating new knowledge. In other words, the emphasis is on (big) data and its intrinsic strength rather than on the information elucidating the context to achieve knowledge. In this scenario, big dataset volume with high variety and accessibility would educate the algorithm and guide AI to the synthesis of these vast observations.

While these statements hold true for some fields such as computer vision, speech recognition, and robotics, another view pictures a more complex scenario, suggesting the development of other areas enhancing human intelligence support and at integration of different AI systems [78].

Use of big data results for developing guidelines

In the last decades, the evidence-based medicine (EBM) movement proposed a straightforward hierarchy of evidence, with emphasis on the ones derived from randomized clinical trials (RCTs) rather than from observational studies [79]. The highest rank of evidence was attributed to systematic reviews, the attempt to evaluate the totality of evidences in order to offer to clinicians the best synthesis of available knowledge. The inclusion in systemic reviews of observational studies, such as cohort and case-control studies, highlights the possibility to derive high level of evidence with study designs other than RCTs [80]. This concept further evolved into the GRADE classification of evidence quality [81], a system that takes into account other elements influencing evidence credibility (i.e., study design, risk of bias, precision, consistency, directness, publication bias, magnitude of effect, and dose-response gradients). This shift empowers observational studies to provide definitive causal evidence [82•].

Clinical practice guidelines embody the attempt to apply available scientific evidences to provide clinicians with guidance in the diagnostic and therapeutic process at the point of care. In doing so, they should be considered a form of decision support system able to standardize the clinical practice, reducing medical errors and inappropriate, inefficient treatments. At times, clinical practice guidelines provide recommendations in the absence of strong evidences. For example, the National Comprehensive Cancer Network (NCCN) clinical guidelines on systemic therapy concomitant to radiotherapy are different for some early-stage oropharyngeal cancers, based on their p16 status [83]. Interestingly, NCCN panel members state that only a small number of patients with such diseases participated in prospective clinical trials, causing the lack of evidence. This is especially true for rare conditions, as in HNCs. In several circumstances, properly designed RCTs or systemic reviews with adequate size to address open clinical issues are simply not feasible. Available examples, such as the meta-analysis on chemo-radiotherapy in HNCs of the MACH-NC Collaborative Group, resulted from large collaborations and provided evidences that still drive daily clinical practice [84, 85]. Unfortunately, this is more of an exception than routine in head and neck oncology. Furthermore, a common critique to EBM-centered clinical practice is the limited population it can be applied to. Given the strict inclusion criteria of clinical trials, a large proportion of the everyday-

population of a head and neck medical oncologist office does not fit in. Nonetheless, one should assume that study-derived evidence would apply to all. But as the MACH-NC meta-analysis showed, the benefit from chemo-radiotherapy rather than from radiotherapy alone in high-risk postoperative setting does not apply to older patients, often excluded from clinical trials.

During the past years, retrospective observational study using population-based cancer registries data evolved, showing the ability to collect clinical information on stage, diagnostic exams, treatments, and follow-up, additionally to the routinely collected cancer registry data. European high-resolution studies conducted on several solid malignancies demonstrated that at least in some European Cancer Registries, this approach is feasible [86], although time and labor consuming. Attempts aiming at data enrichment are ongoing, linking records from cancer registries with different data sources such as health surveys, census, or clinical discharge records [87–90]. However, although these new approaches will provide high amount of data, it still lacks the ability to analyze them with innovative technology in order to provide new insight on clinical issues.

In this scenario, big data-based observational studies could provide the missing piece in the construction of a trustworthy, evidence-based indication for clinical practice. Examples of such approach are already available, showing how innovative data mining leads to novel observations [91]. Therefore, big data-generated evidences should help in driving and sustaining guideline recommendations and reducing the time lag between research and clinical practice. At the same time, big data technology could be used to monitor guideline application or deviation and related clinical outcome, helping a constant updating process and benchmarking activities for quality assurance purpose. To allow this, all factors driving clinician decision to deviate from recommended treatment should be automatically recorded by the system. Since evidence alone never determines decisions, context and values play crucial role in the personalization of health care and development of a reliable decision support system.

Concerns and critics

While big data approach is welcomed by many, others voiced their doubts concerning methodological aspects, unanticipated consequences, and loss of privacy and autonomy [92•, 93•, 94–96]. Medicine has traditionally been a science based on hypotheses which are tested by careful clinical studies [69••, 82•]. On the contrary, practitioners of big data come from a computational tradition that is driven by data rather than by hypothesis testing [75, 92•]. These methods work off raw observations and do not incorporate context knowledge into evidence production. Therefore, an algorithm may detect a pattern in a database but has no way of recognizing whether the result is true, spurious, or affected by bias [69••]. Thus, it is clear why many traditional clinical researchers are questioning the use of big data in health care and research.

Another contradicting point is the amount of data required in order to obtain an accurate, reliable information [95]. Prediction is the most common purpose of big data analysis, as documented by the vast majority of cited papers in the present article designing prognostic algorithms. While modeling techniques for predictions require large datasets [97], forecasting seems to be negatively affected by a merely high amount of data [98].

The adequacy of observational study design based on electronic databases is a cause of concerns. In fact, these investigations pose several challenges that may compromise the validity of such studies [95]. Therefore, methods to minimize the effect of both misclassifications (in the absence of direct assessments of exposure and outcome validity) and confounding elements (in the absence of randomization) need to be implemented [99].

Of note, several independent scientific participants in the ongoing discussion on critical aspects and potential benefits of big data use in medicine agreed upon the need for physicians' training to provide conceptual skills that will enable them to guide this upcoming revolution [76, 77, 100, 101]. In fact, the already ongoing deep interdisciplinary interaction between non-clinical and clinical actors will profoundly shape medical practice evolution toward an "augmented professionalism."

Conclusion

The inevitable technological revolution is already affecting understanding of HNCs, offering opportunities for daily practice improvements. Theory and methodology supporting big data research in health care should be strengthened, especially in the present and upcoming generations of physicians.

Compliance With Ethical Standards

Conflict of Interest

Carlo Resteghini declares that he has no conflict of interest.
Annalisa Trama declares that she has no conflict of interest.
Elio Borgonovi declares that he has no conflict of interest.
Hykel Hosni declares that he has no conflict of interest.
Giovanni Corrao has received research funding through grants from the European Community (EC); the Italian Medicines Agency (AIFA); the Italian Ministry of Education, Universities and Research (MIUR); Novartis; GlaxoSmithKline; Roche; Amgen; and Bristol-Myers Squibb.
Ester Orlandi declares that she has no conflict of interest.
Giuseppina Calareso declares that she has no conflict of interest.
Loris De Cecco has received research funding from the Associazione Italiana Ricerca Cancro (AIRC).
Cesare Piazza declares that he has no conflict of interest.
Luca Mainardi declares that he has no conflict of interest.
Lisa Licitra has received funding (to her institution) for clinical studies and research from AstraZeneca, Boehringer Ingelheim, Eisai, Merck Serono, MSD, Novartis, and Roche; has received compensation for service as a consultant/advisor and/or for lectures from AstraZeneca, Bayer, Bristol-Myers Squibb, Boehringer Ingelheim, Debiopharm, Eisai, Merck Serono, MSD, Novartis, Roche, and Sobi; and has received travel coverage for medical meetings from Bayer, Bristol-Myers Squibb, Debiopharm, Merck Serono, MSD, and Sobi.

Human and Animal Rights and Informed Consent

This article does not contain any studies with human or animal subjects performed by any of the authors.

References and Recommended Reading

Papers of particular interest, published recently, have been highlighted as:

- • Of importance
- •• Of major importance

1. Li S, Kang L, Zhao X-M. A survey on evolutionary algorithm based hybrid intelligence in bioinformatics. *Biomed Res Int*. 2014;2014:1–8. https://doi.org/10.1155/2014/362738.
2. Sessler DI. Big Data–and its contributions to perioperative medicine. *Anaesthesia*. 2014;69(2):100–5 http://www.ncbi.nlm.nih.gov/pubmed/24588022.
3. Baro E, Degoul S, Beuscart R, Chazard E. Toward a literature-driven definition of big data in healthcare. *Biomed Res Int*. 2015;2015:1–9. https://doi.org/10.1155/2015/639021.
4. Frost & Sullivan. Drowning in Big Data? Reducing Information Technology Complexities and Costs For Healthcare Organizations. www.frost.com.
5.• European Commission satellite workshop 'Big data in health research: an EU action plan.' http://bigdata2015.uni.lu/eng/European-Commission-satellite-workshop.

"Big Data in Healthcare" definition for this review was derived by this paper.

6. Fernández-Luque L, Bau T. Health and social media: perfect storm of information. *Healthc Inform Res*. 2015;21(2):67–73. https://doi.org/10.4258/hir.2015.21.2.67.
7. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136(5):E359–86. https://doi.org/10.1002/ijc.29210.
8. Gatta G, Botta L, Sánchez MJ, et al. Prognoses and improvement for head and neck cancers diagnosed in Europe in early 2000s: the EUROCARE-5 population-based study. *Eur J Cancer*. 2015;51(15):2130–43. https://doi.org/10.1016/j.ejca.2015.07.043.
9. Noble WS. What is a support vector machine? *Nat Biotechnol*. 2006;24(12):1565–7. https://doi.org/10.1038/nbt1206-1565.
10. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics*. 2018;15(1):41–51. https://doi.org/10.21873/cgp.20063.
11. Su J, Zhang Y, Su H, Zhang C, Li W. A recurrence model for laryngeal cancer based on SVM and gene function clustering. *Acta Otolaryngol*. 2017;137(5):557–62. https://doi.org/10.1080/00016489.2016.1247984.
12. Yang B, Guo Q, Wang F, Cai K, Bao X, Chu J. A 80-gene set potentially predicts the relapse in laryngeal carcinoma optimized by support vector machine. *Cancer Biomarkers*. 2017;19(1):65–73. https://doi.org/10.3233/CBM-160375.
13. Wan X-B, Zhao Y, Fan X-J, et al. Molecular prognostic prediction for locally advanced nasopharyngeal carcinoma by support vector machine integrated approach. Tao Q, ed. *PLoS One*. 2012;7(3):e31989. doi:https://doi.org/10.1371/journal.pone.0031989
14. Chang S-W, Abdul-Kareem S, Merican A, Zain R. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC Bioinformatics*. 2013;14(1):170. https://doi.org/10.1186/1471-2105-14-170.
15. Li S, Chen X, Liu X, et al. Complex integrated analysis of lncRNAs-miRNAs-mRNAs in oral squamous cell carcinoma. *Oral Oncol*. 2017;73:1–9. https://doi.org/10.1016/j.oraloncology.2017.07.026.
16. Stepp WH, Farquhar D, Sheth S, et al. RNA oncoimmune phenotyping of HPV-positive p16-positive oropharyngeal squamous cell carcinomas by nodal status. *JAMA Otolaryngol Neck Surg*. April 2018. doi:https://doi.org/10.1001/jamaoto.2018.0602
17. Schmidt S, Linge A, Zwanenburg A, et al. Development and validation of a gene signature for patients with head and neck carcinomas treated by postoperative radio(chemo)therapy. *Clin Cancer Res*. 2018;24(6):1364–74. https://doi.org/10.1158/1078-0432.CCR-17-2345.
18. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5(1):4006. https://doi.org/10.1038/ncomms5006.
19. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJWL. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Front Oncol*. 2015;5:272. https://doi.org/10.3389/fonc.2015.00272.
20. Leger S, Zwanenburg A, Pilz K, et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Sci Rep*. 2017;7(1):13206. https://doi.org/10.1038/s41598-017-13448-3.
21.• Parmar C, Leijenaar RTH, Grossmann P, et al. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Sci Rep*. 2015;5(1):11044. https://doi.org/10.1038/srep11044

This paper provided the first insight on radiomic potential in HNCs.

22. Elhalawani H, Kanwar A, Mohamed ASR, et al. Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients. *Sci Rep*. 2018;8(1):1524. https://doi.org/10.1038/s41598-017-14687-0.

23. Elhalawani H, Mohamed ASR, White AL, et al. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. *Sci data*. 2017;4:170077. https://doi.org/10.1038/sdata.2017.77.
24. Ranjbar S, Ning S, Zwart CM, et al. Computed tomography-based texture analysis to determine human papillomavirus status of oropharyngeal squamous cell carcinoma. *J Comput Assist Tomogr*. 2017;42(2):1. https://doi.org/10.1097/RCT.0000000000000682.
25. Lu G, Little JV, Wang X, et al. Detection of head and neck cancer in surgical specimens using quantitative hyperspectral imaging. *Clin Cancer Res*. 2017;23(18):5426–36. https://doi.org/10.1158/1078-0432.CCR-17-0906.
26. Abdollahi H, Mostafaei S, Cheraghi S, Shiri I, Rabi Mahdavi S, Kazemnejad A. Cochlea CT radiomics predicts chemoradiotherapy induced sensorineural hearing loss in head and neck cancer patients: a machine learning and multi-variable modelling study. *Phys Medica*. 2018;45:192–7. https://doi.org/10.1016/j.ejmp.2017.10.008.
27.• Jethanandani A, Lin TA, Volpe S, et al. Exploring applications of radiomics in magnetic resonance imaging of head and neck cancer: a systematic review. *Front Oncol*. 2018;8(MAY):131. https://doi.org/10.3389/fonc.2018.00131.

Usefull review on MRI radiomics in HNCs.

28. Ramkumar S, Ranjbar S, Ning S, et al. MRI-based texture analysis to differentiate sinonasal squamous cell carcinoma from inverted papilloma. *Am J Neuroradiol*. 2017;38(5):1019–25. https://doi.org/10.3174/ajnr.A5106.
29. Zhang B, He X, Ouyang F, et al. Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. *Cancer Lett*. 2017;403:21–7. https://doi.org/10.1016/j.canlet.2017.06.004.
30. Zhang B, Tian J, Dong D, et al. Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. *Clin Cancer Res*. 2017;23(15):4259–69. https://doi.org/10.1158/1078-0432.CCR-16-2910.
31. Wu B, Khong P-L, Chan T. Automatic detection and classification of nasopharyngeal carcinoma on PET/CT with support vector machine. *Int J Comput Assist Radiol Surg*. 2012;7(4):635–46. https://doi.org/10.1007/s11548-011-0669-y.
32. Folkert MR, Setton J, Apte AP, et al. Predictive modeling of outcomes following definitive chemoradiotherapy for oropharyngeal cancer based on FDG-PET image characteristics. *Phys Med Biol*. 2017;62(13):5327–43. https://doi.org/10.1088/1361-6560/aa73cc.
33. Gao RW, Teraphongphom NT, van den Berg NS, et al. Determination of tumor margins with surgical specimen mapping using near-infrared fluorescence. *Cancer Res*. 2018;78(17):5144–54. https://doi.org/10.1158/0008-5472.CAN-18-0878.
34. Farah CS, Fox SA, Dalley AJ. Integrated miRNA-mRNA spatial signature for oral squamous cell carcinoma: a prospective profiling study of narrow band imaging guided resection. *Sci Rep*. 2018;8(1):823. https://doi.org/10.1038/s41598-018-19341-x.
35. Dittberner A, Rodner E, Ortmann W, et al. Automated analysis of confocal laser endomicroscopy images to detect head and neck cancer. *Head Neck*. 2015;38(S1):E1419–26. https://doi.org/10.1002/hed.24253.
36. Mascharak S, Baird BJ, Holsinger FC. Detecting oropharyngeal carcinoma using multispectral, narrow-band imaging and machine learning. *Laryngoscope*. March. 2018. https://doi.org/10.1002/lary.27159.
37. Moccia S, De Momi E, Guarnaschelli M, Savazzi M, Laborai A. Confident texture-based laryngeal tissue classification for early stage diagnosis support. *J Med Imaging*. 2017;4(03):1. https://doi.org/10.1117/1.JMI.4.3.034502.
38. Unger J, Lohscheller J, Reiter M, Eder K, Betz CS, Schuster M. A noninvasive procedure for early-stage discrimination of malignant and precancerous vocal fold lesions based on laryngeal dynamics analysis. *Cancer Res*. 2015;75(1):31–9. https://doi.org/10.1158/0008-5472.CAN-14-1458.
39. Yan B, Li B, Wen Z, Luo X, Xue L, Li L. Label-free blood serum detection by using surface-enhanced Raman spectroscopy and support vector machine for the preoperative diagnosis of parotid gland tumors. *BMC Cancer*. 2015;15(1):650. https://doi.org/10.1186/s12885-015-1653-7.
40. Lau K, Wilkinson J, Moorthy R. A web-based prediction score for head and neck cancer referrals. *Clinical Otolaryngology*. http://www.ncbi.nlm.nih.gov/pubmed/29543399. Published April 6, 2018.
41. Alam IS, Steinberg I, Vermesh O, et al. Emerging intraoperative imaging modalities to improve surgical precision. *Mol Imaging Biol*. June 2018. https://doi.org/10.1007/s11307-018-1227-6.
42. Grillone GA, Wang Z, Krisciunas GP, et al. The color of cancer: margin guidance for oral cancer resection using elastic scattering spectroscopy. *Laryngoscope*. 2017;127:S1–9. https://doi.org/10.1002/lary.26763.
43. Huber T, Wunderling T, Paschold M, Lang H, Kneist W, Hansen C. Highly immersive virtual reality laparoscopy simulation: development and future aspects. *Int J Comput Assist Radiol Surg*. 2018;13(2):281–90. https://doi.org/10.1007/s11548-017-1686-2.
44. Mazur T, Mansour TR, Mugge L, Medhkour A. Virtual reality–based simulators for cranial tumor surgery: a systematic review. *World Neurosurg*. 2018;110:414–22. https://doi.org/10.1016/j.wneu.2017.11.132.
45. Banerjee S, Pal M, Chakrabarty J, et al. Fourier-transform-infrared-spectroscopy based spectral-biomarker selection towards optimum diagnostic differentiation of oral leukoplakia and cancer. *Anal Bioanal Chem*. 2015;407(26):7935–43. https://doi.org/10.1007/s00216-015-8960-3.

46. Liu Y, Li Y, Fu Y, et al. Quantitative prediction of oral cancer risk in patients with oral leukoplakia. *Oncotarget*. 2017;8(28):1–8. https://doi.org/10.18632/oncotarget.17550.
47. Grégoire V. Tumor control probability (TCP) and normal tissue complication probability (NTCP) in head and neck cancer. *Rays*. 30(2):105–8 http://www.ncbi.nlm.nih.gov/pubmed/16294902.
48. Lambin P, van Stiphout RGPM, Starmans MHW, et al. Predicting outcomes in radiation oncology—multifactorial decision support systems. *Nat Rev Clin Oncol*. 2013;10(1):27–40. https://doi.org/10.1038/nrclinonc.2012.196.
49. Dean JA, Wong KH, Welsh LC, et al. Normal tissue complication probability (NTCP) modelling using spatial dose metrics and machine learning methods for severe acute oral mucositis resulting from head and neck radiotherapy. *Radiother Oncol*. 2016;120(1):21–7. https://doi.org/10.1016/j.radonc.2016.05.015.
50. Dean J, Wong K, Gay H, et al. Incorporating spatial dose metrics in machine learning-based normal tissue complication probability (NTCP) models of severe acute dysphagia resulting from head and neck radiotherapy. *Clin Transl Radiat Oncol*. 2018;8:27–39. https://doi.org/10.1016/j.ctro.2017.11.009.
51. Gabryś HS, Buettner F, Sterzing F, Hauswald H, Bangert M. Design and selection of machine learning methods using radiomics and dosiomics for normal tissue complication probability modeling of xerostomia. *Front Oncol*. 2018;8:35. https://doi.org/10.3389/fonc.2018.00035.
52. Pota M, Scalco E, Sanguineti G, et al. Early prediction of radiotherapy-induced parotid shrinkage and toxicity based on CT radiomics and fuzzy classification. *Artif Intell Med*. 2017;81:41–53. https://doi.org/10.1016/j.artmed.2017.03.004.
53. Quon H, Hui X, Cheng Z, et al. Quantitative evaluation of head and neck cancer treatment-related dysphagia in the development of a personalized treatment deintensification paradigm. *Int J Radiat Oncol*. 2017;99(5):1271–8. https://doi.org/10.1016/j.ijrobp.2017.08.004.
54. Zhang HH, D'Souza WD, Shi L, Meyer RR. Modeling plan-related clinical complications using machine learning tools in a multiplan IMRT framework. *Int J Radiat Oncol*. 2009;74(5):1617–26. https://doi.org/10.1016/j.ijrobp.2009.02.065.
55. Guidi G, Maffei N, Vecchi C, et al. A support vector machine tool for adaptive tomotherapy treatments: prediction of head and neck patients criticalities. *Phys Medica*. 2015;31(5):442–51. https://doi.org/10.1016/j.ejmp.2015.04.009.
56. Guidi G, Maffei N, Meduri B, et al. A machine learning tool for re-planning and adaptive RT: a multicenter cohort investigation. *Phys Medica*. 2016;32(12):1659–66. https://doi.org/10.1016/j.ejmp.2016.10.005.
57. Yang X, Wu N, Cheng G, et al. Automated segmentation of the parotid gland based on atlas registration and machine learning: a longitudinal MRI study in head-and-neck radiation therapy. *Int J Radiat Oncol*. 2014;90(5):1225–33. https://doi.org/10.1016/j.ijrobp.2014.08.350.
58. McIntosh C, Welch M, McNiven A, Jaffray DA, Purdie TG. Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method. *Phys Med Biol*. 2017;62(15):5926–44. https://doi.org/10.1088/1361-6560/aa71f8.
59. Niedzwiecki MM, Walker DI, Vermeulen R, Chadeau-Hyam M, Jones DP, Miller GW. The exposome: molecules to populations. *Annu Rev Pharmacol Toxicol*. 2019;59(1):annurev-pharmtox-010818-021315. doi:https://doi.org/10.1146/annurev-pharmtox-010818-021315
60. Irimie AI, Braicu C, Cojocneanu R, et al. Differential effect of smoking on gene expression in head and neck cancer patients. *Int J Environ Res Public Health*. 2018;15(7):1558. https://doi.org/10.3390/ijerph15071558.
61. Gupta S, Tran T, Luo W, et al. Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ Open*. 2014;4(3):e004007. https://doi.org/10.1136/bmjopen-2013-004007.
62. Jiang R, You R, Pei X-Q, et al. Development of a ten-signature classifier using a support vector machine integrated approach to subdivide the M1 stage into M1a and M1b stages of nasopharyngeal carcinoma with synchronous metastases to better predict patients' survival. *Oncotarget*. 2016;7(3):3645–57. https://doi.org/10.18632/oncotarget.6436.
63. Deist TM, Dankers FJWM, Valdes G, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: an empirical comparison of classifiers. *Med Phys*. June. 2018. https://doi.org/10.1002/mp.12967.
64. Van Der Ploeg T, Datema F, Baatenburg De Jong R, Steyerberg EW. Prediction of survival with alternative modeling techniques using pseudo values. Pajewski NM, ed. *PLoS One*. 2014;9(6):e100234. doi:https://doi.org/10.1371/journal.pone.0100234
65. Lan MY, Yang WLR, Lin KT, et al. Using computational strategies to predict potential drugs for nasopharyngeal carcinoma. *Head Neck*. 2014;36(10):1398–407. https://doi.org/10.1002/hed.23464.
66. Randhawa V, Kumar Singh A, Acharya V. A systematic approach to prioritize drug targets using machine learning, a molecular descriptor-based classification model, and high-throughput screening of plant derived molecules: a case study in oral cancer. *Mol Biosyst*. 2015;11(12):3362–77. https://doi.org/10.1039/c5mb00468c.
67. Lewis JS, Ali S, Luo J, Thorstad WL, Madabhushi A. A quantitative histomorphometric classifier (QuHbIC) identifies aggressive versus indolent p16-positive oropharyngeal squamous cell carcinoma. *Am J Surg Pathol*. 2014;38(1):128–37. https://doi.org/10.1097/PAS.0000000000000086.

68.• Big Data and Models for Personalized Head and Neck Cancer Decision Support (BD2Decide). https://clinicaltrials.gov/ct2/show/NCT02832102.

Large internationalretrospective and prospective trial aiming at integration of multiple big data sources to elaborate an HNC decision support system.

69.•• Sim I. Two ways of knowing: big data and evidence-based medicine. *Ann Intern Med*. 2016;164(8):562. https://doi.org/10.7326/M15-2970

This work is of paramount importance for its synthetic and clear dissection of critical points and opportunities in the evolution of scientific and clinical practice.

70. Orlandi E, Licitra L. Personalized medicine and the contradictions and limits of first-generation deescalation trials in patients with human papillomavirus-positive oropharyngeal cancer. *JAMA Otolaryngol Neck Surg*. 2018;144(2):99. https://doi.org/10.1001/jamaoto.2017.2308.

71. Gatta G, Capocaccia R, Botta L, et al. Burden and centralised treatment in Europe of rare tumours: results of RARECAREnet—a population-based study. *Lancet Oncol*. 2017;18(8):1022–39. https://doi.org/10.1016/S1470-2045(17)30445-X.

72.• Goldstein IM, Lawrence J, Miner AS. Human-machine collaboration in cancer and beyond. *JAMA Oncol*. 2017;3(10):1303. https://doi.org/10.1001/jamaoncol.2016.6413

This article drows possible paths for the evolution and integration of "artificial intelligence" in healthcare.

73. Yu K-H, Zhang C, Berry GJ, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun*. 2016;7:12474. https://doi.org/10.1038/ncomms12474.

74. Kohn MS, Sun J, Knoop S, et al. IBM's health analytics and clinical decision support. *IMIA Yearb*. 2014;9(1):154–62. https://doi.org/10.15265/IY-2014-0002.

75. Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics. *JAMA - J Am Med Assoc*. 2018;320(1):E1–2. https://doi.org/10.1001/jama.2018.5602.

76. Licitra L, Trama A, Hosni H. Benefits and risks of machine learning decision support systems. *JAMA*. 2017;318(23):2354. https://doi.org/10.1001/jama.2017.16627.

77. Berner ES, Ozaydin B. Benefits and risks of machine learning decision support systems. *JAMA*. 2017;318(23):2353. https://doi.org/10.1001/jama.2017.16619.

78. Jordan MI. Artificial intelligence — the revolution has not happened yet. https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7. Published 2018.

79. Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA*. 1992;268(17):2420–5.

80. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. *JAMA*. 1995;274(22):1800–4.

81. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. *BMJ*. 2004;328(7454):1490. https://doi.org/10.1136/bmj.328.7454.1490.

82.• Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. *Lancet*. 2017;390(10092):415–23. https://doi.org/10.1016/S0140-6736(16)31592-6

This work reports on the last 25 years of medical evolution.

83. Foote RL, Gilbert J, Gillison ML, et al. *NCCN Guidelines Version 2.2018 Head and Neck Cancers*.; 2018. https://www.nccn.org/professionals/physician_gls/PDF/head-and-neck.pdf.

84. Pignon JP, Bourhis J, Domenge C, Designé L. Chemotherapy added to locoregional treatment for head and neck squamous-cell carcinoma: three meta-analyses of updated individual data. MACH-NC Collaborative Group. Meta-analysis of chemotherapy on head and neck cancer. *Lancet (London, England)*. 2000;355(9208):949–55 http://www.ncbi.nlm.nih.gov/pubmed/10768432.

85. Pignon J-P, le Maître A, Maillard E, Bourhis J, MACH-NC Collaborative Group. Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): An update on 93 randomised trials and 17,346 patients. *Radiother Oncol*. 2009;92(1):4–14. https://doi.org/10.1016/j.radonc.2009.04.014.

86. Trama A, Botta L, Nicolai N, et al. Prostate cancer changes in clinical presentation and treatments in two decades: an Italian population-based study. *Eur J Cancer*. 2016;67:91–8. https://doi.org/10.1016/j.ejca.2016.07.021.

87. Font-gonzalez A, Feijen EL, Sieswerda E, et al. Social outcomes in adult survivors of childhood cancer compared to the general population: linkage of a cohort with population registers. *Psychooncology*. 2015;941(November 2015):933–41. https://doi.org/10.1002/pon.4040.

88. Gunnes MW, Lie RT, Bjørge T, et al. Economic independence in survivors of cancer diagnosed at a young age: a Norwegian national cohort study. *Cancer*. 2016;122(24):3873–82. https://doi.org/10.1002/cncr.30253.

89. Gray L, David Batty G, Craig P, et al. Cohort profile: the Scottish health surveys cohort: linkage of study participants to routinely collected records for mortality, hospital discharge, cancer and offspring birth characteristics in three nationwide studies. *Int J Epidemiol*. 2010;39(2):345–50. https://doi.org/10.1093/ije/dyp155.

90. Leung J, Atherton I, Kyle RG, Hubbard G, McLaughlin D. Psychological distress, optimism and general health in breast cancer survivors: a data linkage study using the Scottish Health Survey. *Support Care Cancer*. 2016;24(4):1755–61. https://doi.org/10.1007/s00520-015-2968-2.

91. Shah NH, LePendu P, Bauer-Mehren A, et al. Proton pump inhibitor usage and the risk of myocardial infarction in the general population. Guo Y, ed. *PLoS One*. 2015;10(6):e0124653. doi:https://doi.org/10.1371/journal.pone.0124653

92.• Coveney PV, Dougherty ER, Highfield RR. Big data need big theory too. *Philos Trans A Math Phys Eng Sci*. 2016;374(2080):20160153. https://doi.org/10.1098/rsta.2016.0153

Critics to the use of big data technology in the absence of adequatetheoretical basis.

93.• Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA*. 2017;318(6):517. https://doi.org/10.1001/jama.2017.7797

Outline of possible pitfall in the use of big data in medicine.

94. Cabitza F, Rasoini R, Gensini GF. Benefits and risks of machine learning decision support systems—reply. *JAMA*. 2017;318(23):2356. https://doi.org/10.1001/jama.2017.16635.

95. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care*. 2013;51(8 SUPPL.3):S30–7. https://doi.org/10.1097/MLR.0b013e31829b1dbd.

96. Auffray C, Balling R, Barroso I, et al. Making sense of big data in health research: towards an EU action plan. *Genome Med*. 2016;8(1):71. https://doi.org/10.1186/s13073-016-0323-y.

97. Van Der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol*. 2014;14(1):137. doi:https://doi.org/10.1186/1471-2288-14-137

98. Hosni H, Vulpiani A. Forecasting in the light of big data. May 2017. doi:https://doi.org/10.1007/s13347-017-0265-3

99. Corrao G. Building reliable evidence from realworld data: methods, cautiousness and recommendations. *Epidemiol Biostat Public Heal*. 2013;10(3):1–40. https://doi.org/10.2427/8981.

100. Lasko TA, Walsh CG, Malin B. Benefits and risks of machine learning decision support systems. *JAMA*. 2017;318(23):2355. https://doi.org/10.1001/jama.2017.16623.

101. Huesch MD. Benefits and risks of machine learning decision support systems. *JAMA*. 2017;318(23):2355. https://doi.org/10.1001/jama.2017.16611.