ELSEVIER

Seminars in
RADIATION
ONCOLOGY

The Evolving Use of Electronic Health Records (EHR) for Research

Ellen Kim, MD, MPH,[*] Samuel M. Rubinstein, MD,[†] Kevin T. Nead, MD, MPhil,[‡] Andrzej P. Wojcieszynski, MD,[‡] Peter E. Gabriel, MD, MSE,[‡] and Jeremy L. Warner, MD, MS[†,§]

Electronic health records (EHR) have been implemented successfully in a majority of United States healthcare systems in some form. There has been a rise in secondary uses of EHR, especially for research. EHR data is large, heterogenous, incomplete, noisy, and primarily created for purposes other than research. This presents many challenges, many of which are beginning to be overcome with the application of computer science artificial intelligence techniques, such as natural language processing and machine learning. EHR are gradually being redesigned to facilitate future research, though we are still far from a "complete EHR."

Semin Radiat Oncol 29:354–361 © 2019 Elsevier Inc. All rights reserved.

Introduction: The Rise of EHR in the United States

Electronic health records (EHRs) are, in the most general sense, clinical information systems that collect, store, and present longitudinal electronic data collected during the delivery of health care.[1] The field of radiation oncology, in particular, has benefited from advancements in treatment planning and delivery systems, which can be considered a special type of EHR. After decades of trying to move from a paper-based record system to an electronic one, attention is now turning away from implementation/adoption and moving toward realizing greater benefits from digital records. This article will focus on the evolving impact of EHRs on research in the United States (US).

The last several decades have seen an expanding role for EHRs. The 1960s and 1970s saw development of the first clinical information systems, such as The Medical Record at Duke University, the Computer Stored Ambulatory Record at Massachusetts General Hospital, and Health Evaluation through Logical Processing at Intermountain Healthcare in Utah.[2] In the early 1970s, the Department of Veterans Affairs developed an EHR, now known as VistA, which remains in use today at the Veterans Health Administration. In the 1980s, the National Academy of Medicine (formerly known as the Institute of Medicine [IOM]) began to analyze paper records and EHRs, and ultimately released reports in 1991 and 1997 that made the case for more widespread EHR adoption in the US healthcare system.[3] In a subsequent study of medical errors, the IOM concluded that computerized order entry was likely to improve patient safety.[3,4] These IOM reports also delineated 8 core functionalities of the EHR (Table 1)[5] and identified barriers to widespread EHR adoption in the US.[3]

Despite this, EHR adoption rates remained low through the 2000s. Although 48% of office-based physicians reported using any EHR-based system in 2009, a survey published that same year indicated that only 1.5% of US hospitals had a "comprehensive" EHR, 7.6% had a basic system, and 17% had computerized provider-order entry.[6,7] Costs of EHR acquisition and maintenance were identified as the major barriers to EHR adoption.[7] In response to this, the HITECH Act, passed as part of the American Recovery and Reinvestment Act of 2009, provided increased federal funding for health IT and established incentives for "Meaningful Use" of EHR technology.[8] President Obama's Precision Medicine Initiative contributed to raising interest in and funding for developing EHRs.[9] Industry-developed systems such as those produced by Epic and Cerner, which focused on optimizing reimbursement, came into widespread use during this time period as well. In response to these and other developments,

[*] Department of Radiation Oncology, Vanderbilt University Medical Center, Nashville, TN

[†] Department of Medicine, Division of Hematology/Oncology, Vanderbilt University Medical Center, Nashville, TN

[‡] Department of Radiation Oncology, University of Pennsylvania School of Medicine, Philadelphia, PA

[§] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

Disclosures: The authors do not have any disclosures or financial relationships with companies.

Address reprint requests to Jeremy Warner, MD, MS, Department of Medicine and Biomedical Informatics, Vanderbilt University Medical Center, 2220 Pierce Ave., 777 PRB, Nashville, TN 37232.
E-mail: Jeremy.warner@vumc.org

https://doi.org/10.1016/j.semradonc.2019.05.010
1053-4296/© 2019 Elsevier Inc. All rights reserved.

Table 1 Core Functionalities of Electronic Health Records (EHR) Defined by the National Academy of Medicine (Formerly Known as the Institute of Medicine [IOM])

	Core Functions	Details
1.	Health information and data	Key information about patients including diagnoses, allergies, medications, lab results, etc. Needs to evolve over time, as new clinical knowledge becomes available and users' needs change.
2.	Result management	Manage results of all types, including lab test results, radiology procedure results, etc. Facilitates efficiency, improves patient safety, reduces redundant testing, aid communication among providers and between providers and patients.
3.	Order entry/ management	Improves workflow processes by eliminating lost orders and ambiguities caused by illegible handwriting, automates creation of related orders, monitors for duplicate orders, reduces the time to fill orders, and improve clinician productivity.
4.	Decision support	Aid providers in disease diagnosis and management. Improve adherence to evidence-based consensus guidelines and protocols.
5.	Electronic communication and connectivity	Facilitate communication among care partners (eg, lab, pharmacy, radiology), enhancing patient safety and quality of care. Improve public health surveillance.
6.	Patient support	Additional support for patient education. May extend to telehealth.
7.	Administrative processes and reporting	Improve efficiency by reducing delays and confusion for patients; improve access to services with immediate validation of insurance eligibility; more timely payments and less paperwork; faster communication of drug recalls; identify candidates for chronic disease management programs.
8.	Reporting and population health management	Represent clinical data with standardized and manipulatable data format to reduce costs associated with reporting key quality indicators.

EHR utilization has increased dramatically – from 2008 to 2015, usage of any type of EHR increased from 9% to 96% for hospitals, and from 17% to 78% for physician offices.[10]

The Primary Intended Purpose of EHR: Billing

As hospitals and ambulatory care centers are increasingly adopting EHRs, most user-level utilization remains focused on billing. The majority of EHRs in the US were designed to support billing,[1] with clinical work flow and now research as secondary applications. In one single-center study, the majority of physicians accessed the EHR to enter data relevant for the purposes of billing (such as clinical documentation), but only a small minority edited problem lists, allergy lists, or suggested changes to the system.[11] Another study demonstrated that, although there is significant heterogeneity in the percentage of providers who use EHR functionalities like clinical decision support tools, and Meaningful Use objective measures that are critical for obtaining optimal reimbursement are fulfilled for virtually all clinical encounters.[12] A significant proportion of the total time and financial cost associated with EHR-based encounters (eg, as high as 25% for emergency department encounters) is related to billing.[13] Diagnosis and procedure billing codes are usually available for clinical encounters, and are thus a robust source of data that can be mined from EHRs for research purposes.

EHR Secondary Purposes

Secondary use of EHRs, including research and teaching, applies to the use of personal health information for uses outside of direct health care delivery.[14] Ironically, this so-called secondary use was the primary reason that medical records were originally kept, which was for didactic purposes.[15] Because research was not the purpose for data creation and the EHR includes sensitive information, research can raise medical ethical, political, technical, and social issues.[16] For example, we recently showed that private patient–provider and provider–provider communications in the EHR can predict for early treatment discontinuation.[17] As the volume of and access to health data increases in both the public and private sectors, it is becoming critical to address these issues. The American Medical Informatics Association (AMIA) published a White Paper[14] describing the need to develop a framework for the secondary use of EHR data, and later a data quality ontology[18] based on Weiskopf's 5 dimensions of data quality.[19]

The secondary applications of EHR are growing as rapidly as the volume of data collected. First, and perhaps most obvious, is their use in more easily automating commonly performed clinical tasks. The use of EHR has made it significantly easier to gather and transfer a vast array of patient data, and the use of EHR may even allow for easier prospective data collection. The retrospective collection of data can be time-consuming, error-prone, and may not be suitable for the purposes of clinical trials. EHR can be linked with electronic health recording instruments in order to more robustly track outcomes such as patient toxicity.[20] For instance, continuous heart rate monitoring could track and more rapidly detect dehydration from poor oral intake of head and neck cancer patients undergoing chemoradiation therapy, leading to more rapid interventions and reducing hospitalizations. The recent announcement that Fitbit data will be included as part of the *All of Us* program's data stream is emblematic of the evolution of such data "mash-ups."[21] There have also been early efforts to capture serial patient-reported outcomes and automatically record them in the EHR.[22]

Although a great deal of data is housed in the EHR, it is in various and often incompatible formats. The most common formats are, in decreasing order of utility: (1) structured and terminology encoded; (2) structured with local codes or no encoding; (3) unstructured machine-readable text; (4) unstructured scanned text (eg, faxes, PDFs). Radiology and pathology reports, for example, are in machine-readable free-text and generally require natural language processing (NLP) approaches to access the data. Although this data would ideally be prospectively coded in a format that could be easily analyzed in the future, this is not the current reality, and automated methods of retrospectively transcoding and extracting data are needed. Systems that are capable of analyzing and robustly extracting clinical data are of significant value, especially in the current era of widely variable standards for data storage, as well as the multiple EHR systems present in most radiation oncology departments (Table 2).[23]

There is no doubt that with the increasing amount of data available to clinicians in the EHR, novel analysis tools may be needed to best make sense of the large amount of data available and to find associations that may not be obvious at first. This is even more apparent in the realm of radiation oncology, where treatment and prognosis not only depend on traditional clinical factors (eg, age, stage, comorbidity, and pathology), but also a wide variety of radiographic and dosimetric data. Furthermore, this data is likely spread across multiple EHR systems for primary records, pathology, radiology, and radiation oncology. The enormous amount of data available in the modern EHR provides a new environment for the application of machine learning (ML) tools and techniques utilizing artificial intelligence. Traditionally, clinicians have been limited to developing nomograms and scoring systems that may only take into account a handful of clinical factors, or be limited in their scope in the interest of ease of use. The use of more sophisticated ML methods may allow for more robust predictive models to be generated using non-traditional variables. These models could ultimately be integrated into the EHR for rapid clinical application.[24-26]

Challenges of Using EHR for Research

There are many challenges of using EHR in research: (1) data quality and validation; (2) complete data capture; (3) heterogeneity among systems; and (4) system knowledge; Cowie et al.[1] describe several examples and potential solutions.

These issues arise because real world data is messy. Many conditions are poorly defined or have conflicting descriptions that are used by different providers, and change over time. For example, there are multiple cancer staging classification systems (eg, American Joint Committee on Cancer, International Federation of Gynecology and Obstetrics, D'Amico risk classification) that have each evolved over time. Even for a seemingly simple lab measurement within one hospital system, there are multiple related data points (eg, orders, billing codes, results) and these are updated at different times in different forms (eg, billing codes, discrete numerical values, PDF, or scanned document). Of course, these data points, formats, and values change over time and the EHR or billing codes are updated. Furthermore, labs have different data than medications, procedures, etc. There are many copies of medical information stored in multiple places being accessed and modified by multiple people simultaneously. Which version of the record reflects the "truth," and how are conflicts resolved? In recognition of the fundamental problem of provenance (who created the data, who recorded the data, who interpreted the data, etc.) the Office of the National Coordinator of Health Information Technology has introduced several "challenges" to improve upon the *status quo*.[27]

There is a standard format called Digital Imaging and Communications in Medicine for storing and transmitting medical imaging data. But there is not yet a widely agreed upon standard format for other EHR data, or even for oncology care. It is notable that the American Society of Clinical Oncology is undertaking an effort to define the Minimal Common Oncology Data Elements to facilitate interoperability and improve data quality for cancer patient care and research.[28] Nevertheless, records are often shared by sending consultation and follow-up notes by mail, fax, or electronic mail. In oncology care, there are times when multiple specialists (radiation oncology, medical oncology, surgical specialties, pathology, radiology) are concurrently caring for the patient from different clinics/hospitals. Even when specialists work within the same health system, radiation oncologists, radiologists, and pathologists often use additional EHR systems that are not seamlessly integrated with the health system's primary EHR, creating additional boundaries to communication.

Overview of EHR Applications for Research

The use of EHR for research has rapidly become ubiquitous. One obvious and unique benefit is the ability to study real-world real-patient outcomes in near-real-time. Analyzing existing data also tends to be less expensive and more convenient than creating a human-curated dataset, whether prospectively or retrospectively. There seems to be a near infinite potential for future applications of EHR in research, especially as the types of collected data and ability to extract information from records continue to improve. For example, in the growing field of radiomics, technology improvements are creating both new MRI sequences and calculating new quantitative features from scans. Already, current applications include observational studies (drug utilization, natural history, risk factors), safety surveillance (post-marketing safety surveillance), regulatory work (safety surveillance, pharmacovigilance)[1]; and clinical research (hypothesis generation, performance improvement, comparative effectiveness[29]).

As the use of EHR for research becomes more popular, EHR are gradually being adapted and redesigned to facilitate research more easily. Attempts are in progress to establish standards in data structure and display; common data elements, such as the North American Association of Central Cancer Registries and STandards for Oncology Registry

Table 2 Approximation of the Current State for the Average Electronic Health Record (EHR)

	Likely to be Found in Structured Data	Likely to be Found in Narrative (+/− Machine Readable)	Unlikely to be Found in the EHR
Facts	Name Identification numbers Date of birth Sex Medications as prescribed	If female: - Number of pregnancies - Age of menarche - Age of menopause Occupation Substance use	Place of birth Sexual history Dietary patterns Living conditions Other social/ behavioural determinants of health Germline DNA or other biologic data Photographs
Observations (provided or elicited from patient)	Race/Ethnicity	Symptoms of illness Family history of illness and/ or longevity	Gender identity Medications as taken
Observations (obtained by the healthcare practitioner and/or system)	Vital signs Most laboratory test results (values)	Signs (other than vital signs) Physical exam findings Some laboratory test results (esoteric labs, somatic DNA testing)	Radiology images Pathology slide images
"Low-level" Interpretations	Some meaning of laboratory test results (normal/ abnormal, high/ low etc.)	Meaning of signs Meaning of physical exam findings Extended meaning of laboratory test results (incl. genomics reports) Radiology reports Pathology reports	
"High-level" Interpretations	Diagnosis (coded)	Diagnosis (descriptive) Risk for developing a diagnosis in the future Prediction (ie, will a future action work or not) Prognosis (ie, what is the expected outcome)	
Plans		Diagnostic workup Lifestyle modifications plan Treatment plan Risk modification plan (eg, preventative measures)	Survivorship care planning
Actions	Starting and modifying medications Removal procedures (eg, excising, eliminating) Additive procedures (eg, device placement) Substitutive procedures (eg, knee replacement) Damaging procedures (eg, radiotherapy)	Observation and monitoring Discontinuing a medication	Lifestyle modifications
Outcomes	Mortality	Success or failure of a plan Treatment summarization Treatment-related adverse events Time to event (eg, PFS) Complications of natural processes	Financial toxicity Patient reported outcomes (PRO) Psychological distress Treatment burden (eg, time spent in appointments)

Entry standards used by national cancer registries; and terminologies for semantic tagging and annotation, such as Systematized Nomenclature of Medicine – Clinical Terms, Logical Observation Identifiers Names and Codes, and RxNorm.[29] But the lack of financial incentives for EHR vendors and users has not facilitated the widespread adoption of standards[29] beyond ICD-9/10 and CPT codes.[30]

Example: Natural Language Processing of EHR Data

The vast majority of EHRs were not designed with the intention of research, so it can be both time-consuming and challenging to convert existing EHR data into a format that can be analyzed. For example, the Surveillance, Epidemiology, and End Results[31] and National Cancer Data Base[32] are national databases built by trained and certified cancer registrars who manually enter relevant clinical information extracted from patients' medical charts. Extracting data in a format that can be analyzed can be time- and cost-intensive.

NLP can convert clinical documents into analyzable data elements, turning big data into smart data.[33] NLP techniques are an application from computer science and computational linguistics, and include processing tasks such as tokenization, named entity recognition, and character gazetteer. For interested readers, there are many tutorials online introducing these concepts, including Coursera, Codecademy, and Universities. Most modern NLP approaches use a combination of rule-based and supervised ML approaches.

NLP has been used to extract cancer stage information[34]; to create oncology treatment summaries[35]; to automate determination of prostate cancer risk group[36]; to categorize oncologic responses in radiology reports[37]; and to mine the EHR for patient-centered outcomes following prostate cancer treatment.[30] While these are specific examples that demonstrate a proof of concept, these NLP techniques have the tremendous benefit of being scalable, adapted, and applied to other similar data. In the future, the role of cancer registrars could be quality control and database maintenance, instead of manual data entry.

Within the past few years, unsupervised methods of NLP, especially word2vec and CUI2vec-based approaches, have become increasingly popular. Essentially, these approaches use context within sentences and other clues in the language itself to make determinations, as opposed to classifying to predetermined labels applied by content experts. In the cancer domain, these techniques have primarily been used in the parsing of pathology reports.[38,39]

Example: Artificial Intelligence and Machine Learning (ML) Using EHR Data

The use of ML algorithms to generate predictive models in medicine has been increasing nearly exponentially. Jiang et al. queried the PubMed database for deep learning in healthcare; the number of articles increased from about 30 in 2013 to nearly 250 in 2016.[40] Supervised, unsupervised, and semisupervised learning algorithms have been used, most commonly using neural networks or support vector machines, and mostly using data from diagnostic imaging.[40]

ML models have already been used in radiation oncology to predict recurrence patterns after intensity modulated radiotherapy (IMRT) for nasopharyngeal cancer[41]; prognosis with glioblastoma[42]; prostate cancer response to IMRT[43,44]; and skin dose in low-kV intraoperative radiotherapy.[45] In medical oncology, ML has been used to predict which patients will acquire a resistance to EGFR inhibitors[46]; to predict breast cancer bone metastasis[47]; and to predict chemotherapy-induced peripheral neuropathy.[48] In radiomics, ML has tried to improve pulmonary nodule screening for lung cancer,[49] to predict mutation status of renal cell carcinoma,[50] and to detect prostate cancer in prostatectomy specimens.[51] In pathology, ML methods have attempted to identify papillary thyroid carcinoma based on the appearance of fine needle aspiration cytology.[52]

Artificial intelligence can be described broadly as the application of computers to mimic the thought processes of human brains; it includes both NLP, ML, and other methods. Artificial intelligence methods can be employed sequentially, using NLP to mine unstructured texts, and ML for creating predictions using the structured output. This sequential strategy has been used to automate extraction of detailed prostate cancer data from clinical notes[53]; to identify local recurrences of breast cancer[54]; and to extract clinical information from discharge summaries.[55]

Regulatory

The US Food and Drug Administration (FDA) has recently expressed increasing interest in using real-world data, primarily derived from EHRs, to inform regulatory decisions, including postmarketing surveillance and decisions about label expansions. In late 2018, FDA commissioner Gottlieb announced the formation of a new office at the FDA, the Framework for the Real-World Evidence Program, dedicated to utilizing this data.[56] The FDA and the pharmaceutical industry have a particular interest in the concept of "synthetic controls" – populations derived from heavily annotated EHR data that can serve as a proxy to a prospective control arm in a randomized controlled trial. Taking this idea further, some have advocated for synthetic controls which are amalgamated across EHRs – the ultimate "average patient." These approaches have generated healthy skepticism but are very likely to inform at least a portion of future regulatory decisions.

Education

As the use of EHR data for research continues to increase in both frequency and complexity, the clinical, informatic, and educator communities will need to collaboratively establish and implement core informatics competencies in healthcare.

A newly offered subspecialty medical certification in Clinical Informatics (https://www.theabpm.org)[57] and a proliferation of high-quality, accessible online education programs, such as the AMIA 10 × 10 initiative (https://www.amia.org/education), represent important training programs in informatics. However, these resources may not reach the average healthcare provider. To combat this, the NIH Big Data to Knowledge (BD2K) invested $200 million to address data science challenges, including educational program development, and HITComp (http://hitcomp.org) has formed as an international effort to standardize health information technology competencies. Ultimately though, the fundamentals of EHR-based research remain unfamiliar to the majority of healthcare professionals, leaving them unprepared to assess the validity and applicability of EHR research to real-world healthcare questions. Multidisciplinary efforts to establish and implement standardized informatics education at all levels of the healthcare training pipeline, such as through the AMIA Informatics Educators Forum (https://www.amia.org/ief2019), are critical to closing the informatics knowledge gap.

In the future, a “Complete EHR”?

In order to anticipate what the future may bring to the EHR, it is useful to look to the past. In 1995, the Computer-based Patient Record Institute Work Group defined the EHR as “a virtual compilation of nonredundant health data about a person across a lifetime, including facts, observations, interpretations, plans, actions, and outcomes.”[3] While this definition is laudable and provides one with a useful framework, no current EHR technology meets this extremely broad definition, nor is this a goal of current EHR technology. In fact, EHR data is highly redundant and nonstandardized.[58] Some facts, observations, and actions are captured in a structured fashion, whereas the majority of interpretations, plans, and outcomes are represented in an unstructured fashion, when they are present at all in the EHR.

Another useful framework is the concept of the Designated Record Set.[59] Per the American Health Information Management Association, “there is no one-size-fits-all definition for the designated record set. The healthcare organization must explicitly define both in a multidisciplinary team approach. Medical staff, for example, should provide guidance to ensure that patient care needs will be met for immediate, long-term, and research uses.” Unfortunately, even if it were well defined, the Designated Record Set is more geared to describing the current, as opposed to the possible. Many clinical data elements, some of which may be critically important to the diagnosis, treatment, and continuing care of patients, are simply not present in the EHR with any predictability. Much of what has been coined the “health tapestry” by Weber et al. is completely absent or sparsely recorded within clinical notes.[60]

Using the Computer-based Patient Record Institute model, we have approximated the current state (as of 2019) for the average “comprehensive” EHR,[7] in Table 1. For each of the categories of facts, observations, interpretations, plans, actions, and outcomes, we have determined: (1) which data elements are likely to be found in structured data; (2) which are likely to be found in free text (some of which will be amenable to extraction through NLP); and (3) which are unlikely to be found in the EHR. As we anticipate which data streams EHR will begin to capture in the future, it is possible that we may also witness revolutions in clinical documentation, such as the introduction of the SOAP note in the 1960s[61] and our recent proposal to “wikify” the medical record.[62]

Conclusions

In recent decades, EHRs have been adopted in nearly every healthcare system in the US. Due to the financial incentives, EHR have been traditionally designed to aid administration (registration, scheduling, billing) and basic clinical care, rather than research. With the inflation-adjusted decline in research funding, rise in costs of traditional gold-standard prospective randomized clinical trials, and rise in computer science applications for healthcare data, it seems obvious to try to harness this growing amount of data to improve clinical outcomes and efficiency. EHR are gradually starting to be re-designed to be more useful for research purposes, and we are finding new techniques for overcoming limitations. The use of EHR for research is an exciting field with dynamic advances to constantly evolving data and has amazing potential.

References

1. Cowie MR, Blomster JI, Curtis LH, et al: Electronic health records to facilitate clinical research. Clin Res Cardiol 106:1-9, 2017. https://doi.org/10.1007/s00392-016-1025-6
2. Lloyd SC: Clinical decision support systems for ambulatory care. Proc Annu Sympos Comput Appl Med Care 1984: 470-475. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2578687/
3. Institute of Medicine (US) Committee on Improving the Patient Record. The Computer-Based Patient Record: Revised Edition: An Essential Technology for Health Care. In: Dick RS, Steen EB, Detmer DE, eds. The Computer-Based Patient Record: Revised Edition: An Essential Technology for Health Care, Washington (DC): National Academies Press (US); 1997. http://www.ncbi.nlm.nih.gov/books/NBK233047/. Accessed January 18, 2019
4. Institute of Medicine (US) Committee on Quality of Health Care in America. To Err Is Human: Building a Safer Health System. In: Kohn LT, Corrigan JM, Donaldson MS, eds. To Err Is Human: Building a Safer Health System, Washington (DC): National Academies Press (US); 2000. http://www.ncbi.nlm.nih.gov/books/NBK225182/. Accessed January 18, 2019
5. Institute of Medicine (US) Committee on Data Standards for Patient Safety. Key Capabilities of an Electronic Health Record System: Letter Report. Washington (DC): National Academies Press (US), 2003. http://www.ncbi.nlm.nih.gov/books/NBK221802/. Accessed February 2, 2019
6. Hsiao C-J, Hing E, Ashman J: Trends in electronic health record system use among office-based physicians: United States, 2007−2012. Natl Health Stat Rep 2014: 1-18
7. Jha AK, DesRoches CM, Campbell EG, et al: Use of electronic health records in US hospitals. N Engl J Med 360:1628-1638, 2009. https://doi.org/10.1056/NEJMsa0900592
8. Blumenthal D, Tavenner M: The “meaningful use” regulation for electronic health records. N Engl J Med 363:501-504, 2010. https://doi.org/10.1056/NEJMp1006114
9. Yim W-W, Wheeler AJ, Curtin C, et al: Secondary use of electronic medical records for clinical research: Challenges and opportunities.

Convergent Sci Phys Oncol 4: 2018.:014001. https://doi.org/10.1088/2057-1739/aaa905

10. 2016 Report To Congress on Health IT Progress. Washington (DC): Office of the National Coordinator for Health Information Technology (ONC) Office of the Secretary, United States Department of Health and Human Services, 2016. p 32. https://www.healthit.gov/sites/default/files/2016_report_to_congress_on_healthit_progress.pdf. Accessed January 18, 2019.
11. Wilcox A, Bowes WA, Thornton SN, et al: Physician use of outpatient electronic health records to improve care. AMIA Annu Sympos Proc 2008:809-813, 2008
12. Ancker JS, Kern LM, Edwards A, et al: How is the electronic health record being used? Use of EHR data to assess physician-level variability in technology use. J Am Med Inf Assoc 21:1001-1008, 2014. https://doi.org/10.1136/amiajnl-2013-002627
13. Tseng P, Kaplan RS, Richman BD, et al: Administrative costs associated with physician billing and insurance-related activities at an Academic Health Care System. JAMA 319:691-697, 2018. https://doi.org/10.1001/jama.2017.19148
14. Safran C, Bloomrosen M, Hammond WE, et al: Toward a national framework for the secondary use of health data: An American Medical Informatics Association White Paper. J Am Med Inf Assoc 14:1-9, 2007. https://doi.org/10.1197/jamia.M2273
15. Gillum RF: From papyrus to the electronic tablet: A brief history of the clinical medical record with lessons for the digital age. Am J Med 126:853-857, 2013. https://doi.org/10.1016/j.amjmed.2013.03.024
16. Miller ST, Pickering RG: Use of electronic patient data in research. AMA J Ethics 13:148-151, 2011. https://doi.org/10.1001/virtualmentor.2011.13.3.ccas2-1103
17. Yin Z, Harrell M, Warner JL, et al: The therapy is making me sick: How online portal communications between breast cancer patients and physicians indicate medication discontinuation. J Am Med Inf Assoc 25:1444-1451, 2018. https://doi.org/10.1093/jamia/ocy118
18. Johnson SG, Speedie S, Simon G, et al: A data quality ontology for the secondary use of EHR data. AMIA Annu Sympos Proc 2015:1937-1946, 2015
19. Weiskopf NG, Weng C: Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. J Am Med Inf Assoc 20:144-151, 2013. https://doi.org/10.1136/amiajnl-2011-000681
20. Albuquerque K, Rodgers K, Spangler A, et al: Electronic medical record-based radiation oncology toxicity recording instrument aids benchmarking and quality improvement in the clinic. J Oncol Pract 14:e186-e193, 2018. https://doi.org/10.1200/JOP.2017.025163
21. National Institute of Health (NIH). All of US research program expands data collection efforts with Fitbit. https://allofus.nih.gov/news-events-and-media/announcements/all-us-research-program-expands-data-collection-efforts-fitbit. Published January 16, 2019. Accessed 2 February, 2019
22. Stehlik J, Rodriguez-Correa C, Spertus JA, et al: Implementation of real-time assessment of patient-reported outcomes in a heart failure clinic: A feasibility study. J Card Fail 23:813-816, 2017. https://doi.org/10.1016/j.cardfail.2017.09.009
23. Khor RC, Nguyen A, O'Dwyer J, et al: Extracting tumour prognostic factors from a diverse electronic record dataset in genito-urinary oncology. Int J Med Inf 121:53-57, 2019. https://doi.org/10.1016/j.ijmedinf.2018.10.008
24. Gensheimer MF, Henry AS, Wood DJ, et al: Automated survival prediction in metastatic cancer patients using high-dimensional electronic medical record data. J Natl Cancer Inst. October 2018. https://doi.org/10.1093/jnci/djy178
25. Lindsay WD, Ahern CA, Tobias JS, et al: Automated data extraction and ensemble methods for predictive modeling of breast cancer outcomes after radiation therapy. Med Phys November 2018. https://doi.org/10.1002/mp.13314
26. Nakatsugawa M, Cheng Z, Kiess A, et al: The needs and benefits of continuous model updates on the accuracy of RT-induced toxicity prediction models within a learning health system. Int J Radiat Oncol Biol Phys October 2018. https://doi.org/10.1016/j.ijrobp.2018.09.038
27. Provenance Challenge. *CCC Innovation Center*. https://www.cccinnovationcenter.com/challenges/provenance-challenge/. Accessed February 3, 2019.
28. American Society of Clinical Oncology (ASCO). View and Comment on ASCO's mCODETM Data Specification First Draft. ASCO. https://www.asco.org/advocacy-policy/asco-in-action/view-and-comment-on-mcode-data-specification-first-draft. Published January 22, 2019. Accessed February 2, 2019
29. Manion FJ, Harris MR, Buyuktur AG, et al: Leveraging EHR data for outcomes and comparative effectiveness research in oncology. Curr Oncol Rep 14:494-501, 2012. https://doi.org/10.1007/s11912-012-0272-6
30. Hernandez-Boussard T, Kourdis PD, Seto T, et al: Mining electronic health records to extract patient-centered outcomes following prostate cancer treatment. AMIA Annu Sympos Proc 2017:876-882, 2018
31. National Cancer Institute: SEER Surveillance, Epidemiology, and End Results. SEER. https://seer.cancer.gov/index.html. Accessed January 12, 2019
32. Bilimoria KY, Stewart AK, Winchester DP, et al: The national cancer data base: A powerful initiative to improve cancer care in the United States. Ann Surg Oncol 15:683-690, 2008. https://doi.org/10.1245/s10434-007-9747-3
33. HealthITSecurity. The difference between big data and smart data in healthcare. HealthIT Anal. https://healthitanalytics.com/features/the-difference-between-big-data-and-smart-data-in-healthcare. Published June 15, 2016. Accessed January 13, 2019.
34. Warner JL, Levy MA, Neuss MN, et al: ReCAP: Feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. J Oncol Pract 12:157-158; e169−167, 2016. https://doi.org/10.1200/JOP.2015.004622
35. Warner JL, Anick P, Hong P, et al: Natural language processing and the oncologic history: Is there a match? J Oncol Pract 7:e15-e19, 2011. https://doi.org/10.1200/JOP.2011.000240
36. Gregg JR, Lang M, Wang LL, et al: Automating the determination of prostate cancer risk strata from electronic medical records. JCO Clin Cancer Inf 2017: 2017. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5847303/. Accessed January 13, 2019
37. Chen P-H, Zafar H, Galperin-Aizenberg M, et al: Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. J Digit Imaging 31:178-184, 2018. https://doi.org/10.1007/s10278-017-0027-x
38. Gao S, Young MT, Qiu JX, et al: Hierarchical attention networks for information extraction from cancer pathology reports. J Am Med Inf Assoc November 2017. https://doi.org/10.1093/jamia/ocx131
39. Yala A, Barzilay R, Salama L, et al: Using machine learning to parse breast pathology reports. Breast Cancer Res Treat 161:203-211, 2017. https://doi.org/10.1007/s10549-016-4035-1
40. Jiang F, Jiang Y, Zhi H, et al: Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2:230-243, 2017. https://doi.org/10.1136/svn-2017-000101
41. Li S, Wang K, Hou Z, et al: Use of Radiomics combined with machine learning method in the recurrence patterns after intensity-modulated radiotherapy for nasopharyngeal carcinoma: A preliminary study. Front Oncol 8:648, 2018. https://doi.org/10.3389/fonc.2018.00648
42. Wong KK, Rostomily R, Wong STC: Prognostic gene discovery in glioblastoma patients using deep learning. Cancers (Basel) 11: 2019.. https://doi.org/10.3390/cancers11010053
43. Abdollahi H, Moid B, Shiri I, et al: Machine learning-based radiomic models to predict intensity-modulated radiation therapy response, Gleason score and stage in prostate cancer. Radiol Med January 2019. https://doi.org/10.1007/s11547-018-0966-4
44. Peeken JC, Goldberg T, Pyka T, et al: Combining multimodal imaging and treatment features improves machine learning-based prognostic assessment in patients with glioblastoma multiforme. Cancer Med December 2018. https://doi.org/10.1002/cam4.1908
45. Avanzo M, Pirrone G, Mileto M, et al: Prediction of skin dose in low-kV intraoperative radiotherapy using machine learning models trained on results of in vivo dosimetry. Med Phys January 2019. https://doi.org/10.1002/mp.13379

46. Kim YR, Kim YW, Lee SE, et al: Personalized prediction of acquired resistance to EGFR-targeted inhibitors using a pathway-based machine learning approach. Cancers (Basel) 11: 2019.(1). https://doi.org/10.3390/cancers11010045
47. Park SB, Chung CK, Gonzalez E, et al: Causal inference network of genes related with bone metastasis of breast cancer and osteoblasts using causal Bayesian networks. J Bone Metab 25:251-266, 2018. https://doi.org/10.11005/jbm.2018.25.4.251
48. Bloomingdale P, Mager DE: Machine learning models for the prediction of chemotherapy-induced peripheral neuropathy. Pharm Res 36:35, 2019. https://doi.org/10.1007/s11095-018-2562-7
49. Alahmari SS, Cherezov D, Goldgof D, et al: Delta radiomics improves pulmonary nodule malignancy prediction in lung cancer screening. IEEE Access 6:77796-77806, 2018. https://doi.org/10.1109/ACCESS.2018.2884126
50. Kocak B, Durmaz ES, Ates E, et al: 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 January 2019: 1-9. https://doi.org/10.2214/AJR.18.20443
51. Aubertin K, Desroches J, Jermyn M, et al: Combining high wavenumber and fingerprint Raman spectroscopy for the detection of prostate cancer during radical prostatectomy. Biomed Opt Express 9:4294-4305, 2018. https://doi.org/10.1364/BOE.9.004294
52. Sanyal P, Mukherjee T, Barui S, et al: Artificial intelligence in cytopathology: A neural network to identify papillary carcinoma on thyroid fine-needle aspiration cytology smears. J Pathol Inf 9:43, 2018. https://doi.org/10.4103/jpi.jpi_43_18
53. Leyh-Bannurah S-R, Tian Z, Karakiewicz PI, et al: Deep learning for natural language processing in urology: State-of-the-art automated extraction of detailed pathologic prostate cancer data from narratively written electronic health records. JCO Clin Cancer Inf 2018: 1-9. https://doi.org/10.1200/CCI.18.00080
54. Zeng Z, Espino S, Roy A, et al: Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinf 19(Suppl 17):498, 2018. https://doi.org/10.1186/s12859-018-2466-x
55. Jiang M, Chen Y, Liu M, et al: A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inf Assoc 18:601-606, 2011. https://doi.org/10.1136/amiajnl-2011-000163
56. Gottlieb S. Statement from FDA Commissioner Scott Gottlieb, M.D., on FDA's New Strategic Framework to Advance Use of Real-World Evidence to Support Development of Drugs and Biologics. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm627760.htm. Accessed February 2, 2019.
57. Detmer DE, Munger BS, Lehmann CU: Clinical informatics board certification: History, current status, and predicted impact on the clinical informatics workforce. Appl Clin Inf 1:11-18, 2010. https://doi.org/10.4338/ACI-2009-11-R-0016
58. Wrenn JO, Stein DM, Bakken S, et al: Quantifying clinical narrative redundancy in an electronic health record. J Am Med Inf Assoc 17:49-53, 2010. https://doi.org/10.1197/jamia.M3390
59. Haugen MB, Tegen A, Warner D, et al: Fundamentals of the legal health record and designated record set. J AHIMA 82:44-49, 2011
60. Weber GM, Mandl KD, Kohane IS: Finding the missing link for big biomedical data. JAMA 311:2479-2480, 2014. https://doi.org/10.1001/jama.2014.4228
61. Weed LL: Medical records that guide and teach. N Engl J Med 278:593-600, 1968. https://doi.org/10.1056/NEJM196803142781105
62. Warner JL, Smith J, Wright A: It's time to wikify clinical documentation: How collaborative authorship can reduce the burden and improve the quality of the electronic health record. Acad Med January 2019. https://doi.org/10.1097/ACM.0000000000002613