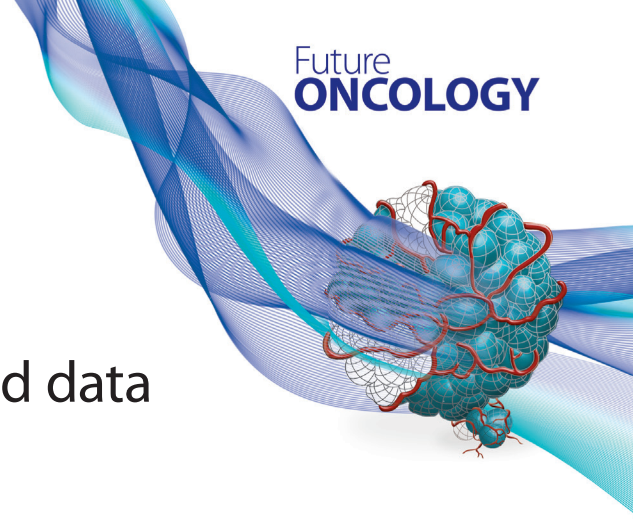
Opportunities and Challenges in Leveraging Electronic Health Record Data in Oncology

Article *in* Future Oncology · March 2016

DOI: 10.2217/fon-2015-0043



Future
ONCOLOGY

PERSPECTIVE

For reprint orders, please contact: reprints@futuremedicine.com

Opportunities and challenges in leveraging electronic health record data in oncology

Marc L Berger*[,1], Melissa D Curtis[2], Gregory Smith[1], James Harnett[1] & Amy P Abernethy[2]

The widespread adoption of electronic health records (EHRs) and the growing wealth of digitized information sources about patients is ushering in an era of ‘Big Data’ that may revolutionize clinical research in oncology. Research will likely be more efficient and potentially more accurate than the current gold standard of manual chart review studies. However, EHRs as they exist today have significant limitations: important data elements are missing or are only captured in free text or PDF documents. Using two case studies, we illustrate the challenges of leveraging the data that are routinely collected by the healthcare system in EHRs (e.g., real-world data), specific challenges encountered in the cancer domain and opportunities that can be achieved when these are overcome.

First draft submitted: 1 December 2015; Accepted for publication: 1 February 2016; Published online: 8 March 2016

KEYWORDS
- clinical research
- electronic health records
- real-world data

Three major trends are converging to revolutionize the ability of aggregate data to impact the management of patients with cancer. First, genomics and other so-called ‘omics’ (metabolomic, proteomic and microbiomic) are transforming our understanding of the pathogenesis of cancer. Second, the widespread adoption of electronic health records (EHRs) and the growing wealth of digitized information sources about patients, including patient-reported outcomes, data streams from wearable sensors and their linkage to ‘omic’ data, have ushered in an era of ‘Big Data’ for healthcare [1]. Finally, the frontiers of analysis are rapidly expanding to deal with this increase in data. Advanced analytic approaches, including machine learning and new statistical methods, promise to uncover previously unidentified associations and accelerate knowledge discovery.

The convergence of these trends opens multiple avenues to advance the field of cancer research and treatment, including such examples as retrospective assessment of known findings in clinical practice settings, discovery of novel associations in the context of gene–environment interactions and prospective application to decision support. Mining of EHR data also holds the promise of providing information on larger populations of cancer patients to understand the pharmacoepidemiology, pharmacoeconomics and patient experience of cancer and cancer therapy. To the extent that these data are found to be accurate enough and useful, the evolving landscape of Big Data will be a powerful addition to available cancer registries and databases. Newer data resources are also likely to be much more efficient, and potentially more accurate, than the current gold standard: chart review studies to extract all desired data elements.

That being said, the promise of the Big Data era in healthcare has been tempered by the reality that it fundamentally builds from a backbone of EHR data, and EHR data suffers from quality

[1]Pfizer Inc., 235 East 42nd Street, New York, NY 10017, USA
[2]Flatiron Health, 96 Spring Street, New York, NY 10012, USA
*Author for correspondence: marc.berger@pfizer.com

Future Medicine part of fsg

10.2217/fon-2015-0043 © 2016 Future Medicine Ltd ISSN 1479-6694

issues – inconsistent formatting, missingness, variability in recording and/or many important variables are not routinely coded but contained in free text. Digitized EHR data (also known as 'structured data', i.e., the information in the EHR that is standardly coded such as height/weight, gender, centralized laboratory tests, medication orders) can be a rich source but it needs to be combined into a common analyzable format and still only tells part of the clinical story. 'Unstructured data' (i.e., data in the EHR that is not entered directly into a discrete field, but rather is brought in through other documentation such as pathology reports, radiology reports and physician notes, among others) can help to complete the dataset but these data are difficult to extract in an accurate, reliable and cost-effective fashion. Standard 'chart review' approaches are slow, expensive and often lack quality control [2]. Publically mandated tumor registries are often 6–24 months behind in coding [3,4], and focus on a sparse set of variables predominantly needed to support epidemiological research initiatives. Natural language processing (NLP) provides promising tools to make sense of this unstructured data, however the state of the art techniques do not yet by themselves give human-level accuracy. The unstructured data problem is a major limitation of EHR data, and solving it would represent a major advancement in our ability to use data to effect meaningful change.

There are numerous efforts underway to create infrastructure to efficiently collect, analyze and derive value from these new data sources. Initiatives such as CancerLinQ, IBM Watson, the Oncology Research Information Exchange Network and Flatiron Health all represent important initiatives to address the unstructured data challenge in oncology. This article will focus on the challenges of leveraging the data that are routinely collected by the healthcare system in EHRs (e.g., real-world data), specific challenges encountered in the cancer domain and results that can be achieved when these problems are overcome.

General challenges of EHRs

In the USA, EHRs have only been in mainstream use for about the last 20 years. During this latter period, EHRs were primarily designed to assist with workflow and billing. As a result, they were not initially designed to provide clinical decision support nor were they designed to facilitate health services research. With the incentives provided by the HITECH Act in 2009, adoption of EHRs has moved beyond leading medical centers to their widespread use by most hospitals and larger physician practices. This revolution has not been without its downsides [5] and the disruption has led to substantial dissatisfaction by practicing physicians [6].

Healthcare providers are not finding immediate value in rigorously completing data entry in the electronic chart – it frequently does not help them provide care more efficiently to their patients. Doctors cite more time in front of a computer, less time with patients and clinically-unnecessary alert fatigue; indeed, when decision support reminders are provided, they are often ignored or turned off [7,8]. Furthermore, the practice of medicine is more of an evolving narrative than structured events, and doctors find it hard to reduce clinical complexity into a series of checkboxes [9]. Thus, it is not surprising that the first challenge in leveraging real world data from EHRs is that structured data entry in EHRs is far from complete.

The second challenge is that information technology systems are highly fragmented both across and within healthcare organizations [10,11]. Between organizations, the data systems among providers are poorly connected, and, currently, there is little incentive to solve this. To date, EHR vendors have little to gain from making their offerings interoperable, although recent evolutions in the regulatory landscape (e.g., meaningful use) as well as technology solutions like application program interfaces may shift this dynamic moving forward. Moreover, many enterprise EHR installations are customized such that even installations from the same vendor cannot readily communicate with installations at other institutions. Patient and clinician demand are helping to improve this, and as a result, interoperability has been identified as a critical obstacle to overcome in realizing the vision of a learning healthcare system. This issue was recently highlighted when the American Society of Clinical Oncology issued a position statement calling for legislation to mandate the interoperability of EHRs [12]. Progress in this arena will depend on addressing misaligned incentives among various components of the healthcare ecosystem.

The third challenge is that the majority of data is recorded as free text [13] or is trapped in scanned documents (e.g., PDF images of radiology reports). Buried in these notes are critical elements of the patient experience (e.g., reason for treatment discontinuation, toxicities and functional status, among others). Interestingly, the proportion of critical data in unstructured documents is increasing, not decreasing, fueled by doctors needing to document in an efficient manner, increasing scrutiny via recovery audit contractor audits, among others. There are many ongoing efforts to use NLP to create standardized coded information from free text notes [14], but to date, such machine-learning techniques have not led to adequate high-quality data that can be used for research or clinical purposes in cancer; we note that there are some compelling examples of NLP being useful outside of the cancer setting [15–19]. Evolving methods focus on combinations of human review (e.g., updated approaches that mimic chart review or tumor registries) plus technology interfaces to start to accommodate the 'unstructured data' problem. There is also a need for advances in normative terminology, ontologies and machine-learning approaches.

Specific challenges with oncology EHRs

To understand the treatment of oncology patients, it is critical that there be complete information for a variety of data points including biomarkers, disease staging, location of metastases, disease progression, among others. More than half of these critical variables are only recorded in unstructured data [20]. Critical data points require identification and interpretation in order to generate them and make them analyzable. Data elements like cancer stage can only be generated after carefully reviewing myriad documents and assembling the data point. At the current time, this requires trained staff. Either the oncology care team needs to do it up front, or certified abstractors and cancer tumor registrars need to do it on the back end. This makes the problem of data mining in oncology more challenging than it is in other diseases where, for example, the biomarkers are routinely captured in laboratory data (e.g., LDL-C, Hgb A1C), or where disease staging and progression can be inferred from a combination of insurance claims and EHR data. The process of chart abstraction can be technology enabled, which can enhance the efficiency of the process. Advancements in NLP and cognitive computing (e.g., IBM Watson) may make the process less onerous in the future. However, as it stands today, the process of cognitive computing requires significant and ongoing input from clinical and scientific experts.

Achieving the vision of Big Data for oncology

Despite these challenges, there is hope for the future of EHR data in oncology and its potential to support the Big Data vision. Industry, professional bodies, philanthropic organizations, government agencies and physician groups are all working to solve the challenges outlined, as well as corollary problems such as supporting patient privacy, accommodating regulatory requirements and data linkage. For example, in the private sector, some companies have created successful business models that enable the linkage of data by providing dashboards and reports to provider organizations that facilitate population and patient management as well as fulfill requirements for quality standards; they also offer value to providers as they become part of a larger network with the common goal of improving patient outcomes. The byproduct is the creation of integrated datasets (within a common data framework) that facilitate health services and outcomes research. In another example, companies are working on various technologies to mine unstructured documents and digitize critical data elements; some of these involve people reviewing documents by hand but at scale and others are electronic only.

As larger digitized datasets become available, populated using data from the EHR, the promise of Big Data in oncology becomes more possible. Genomic, claims, patient-reported and other data can be linked to the EHR dataset. The EHR forms a longitudinal digital story of the patient experience – an analytic backbone, with which all of the additional data points can be temporally linked. Standardized and novel analyses can run on top of this dataset, providing access to new insights at scale. In the next section, we explore two case studies that demonstrate the EHR data challenges that must be solved, and the potential implications and possible uses for the data when they are solved.

Practical applications: cohort selection is the most critical challenge in analyzing EHR data

In using EHR data, cohort selection is the most fundamental analytic task, because inaccurate cohorts lead to inaccurate results. To illustrate this, we will discuss two case studies, both related to non-small-cell lung cancer (NSCLC). Identifying the patient cohort for analysis is complicated by the fact that all of lung cancer relies on the same ICD-9 code even though small cell lung cancer is different from NSCLC in terms of clinical course and treatment. Further, squamous NSCLC is a unique histology and has a poorer prognosis; these patients are more likely to be smokers without a targeted *EGFR* mutation. To compare outcomes with chemotherapy, the relevant population are NSCLC patients with nonresectable disease (so called 'advanced'), which includes some stage IIIB, as well as IV and recurrent cases; including only patients with metastatic disease would miss critical patients receiving treatment. Finally, in order to examine survival outcomes, EHR data must be linked to other datasets that incorporate the outcomes of interest. For example, assessments of treatment impact on longevity require linkage to mortality data, which in itself can be challenging due to the limitations associated with available sources [21,22]. An assessment of cost and value would require access to linked closed claims data or another similar source.

Historically, developing analytic cohorts in a disease of interest was accomplished using administrative data like Medicare or Surveillance, Epidemiology and End Result Program (SEER) data. The ICD-9 code for lung cancer would be used to screen for the population of interest. Since the ICD-9 code (and ICD-10 code) does not differentiate between small-cell (SCLC) and NSCLC types of lung cancer, a proxy might be applied to filter out inappropriate patients. For example, administration of etoposide might be used to exclude SCLC patients from the analytic cohort; however, etoposide is the seventh most commonly used chemotherapy in NSCLC, so that some of the NSCLC cohort would be erroneously excluded. In order to find people suffering from 'advanced disease' (IIIB, IV or recurrent), the only ICD-9 code option is one for secondary metastases. However, this code does not discriminate based upon the etiology of the metastasis, so a person with an early stage lung cancer and metastatic prostate cancer would erroneously end up in the study sample. In addition, the ICD-9 codes for secondary metastases are only used approximately 20–40% of the time in advanced cancer so most of the truly 'advanced' patients will be missing from the analytic cohort. Finally, there is no way to use ICD-9 (or ICD-10) codes to find a person with squamous cell histology, so analyses focused on this important subgroup would not be possible in this framework.

A better approach would be to pull the potential sample of patients based upon the lung cancer ICD-9 code, then look into the medical chart documents and confirm key features of the disease such as squamous cell histology and advanced disease status. A 'pure cohort' can be generated in this way and comparative effectiveness analyses and other inquiries can be executed. Unfortunately, these critical clinical details are not present in the chart in a digitized format – they are hidden in medical case notes, radiology reports and pathology reports as PDF documents. Unstructured data processing of some type (e.g., chart review, technology-enabled abstraction) is required to accomplish even this basic early task of observational cancer research using EHR data – cohort selection.

The following call-out boxes demonstrate [1]: how combining structured and unstructured EHR data contribute meaningfully to analyses in the Big Data landscape [2]; the importance of using accurate cohort selection for subsequent analyses; and [3] the power of data linkage within the EHR-derived Big Data context. The first example compares the results of an analysis generated using only structured EHR data to one where both structured and unstructured data was used to select the correct patient cohort, including an assessment of mortality. The second example demonstrates the importance of complete data capture for patient characteristics and outcomes (including mortality and healthcare resource use) that require integration of claims and other data sources to assess the value of an intervention.

• Case study: assessing the management of NSCLC cancer in the community with Flatiron technology

To demonstrate the impact of accurate cohort selection on study results, we compared the results of an analysis using two cohorts, one based on structured data exclusively and one based on the combination of structured and unstructured data. Patients were identified using a large EHRs database that includes data on more than 850,000 cancer patients treated at approximately 200 distinct community oncology practices geographically distributed throughout the USA (Flatiron Health, NY, USA, data through 30 September 2015). Using Flatiron's approach, a full copy of each patient's medical record is pulled into a central repository for processing. Structured data such as demographics, medications and routine laboratory tests are harmonized and normalized to a standard ontology and common data model (e.g., laboratory data are mapped to LOINC). These structured data are processed and harmonized centrally by Flatiron's technology-assisted data engine and made accessible for research and analytics. Unstructured data such as PDF files representing clinical case notes, pathology reports, radiology reports and complex laboratory tests (e.g., next-generation sequencing) are turned into discrete analyzable data using technology-enabled abstraction [23]. This method combines trained human abstractors with software specifically developed for the identification and targeted display of selected portions of the patient chart. This approach allows information to be captured uniformly, quality to be continuously measured and both structured and unstructured EMR data to be used for analyses. The modular and targeted approach allows chart abstractors to focus their attention on specific data elements, leading to faster, scalable and more accurate data collection.

The two cohorts generated from Flatiron data are shown in Table 1, namely structured EHR data only ('Structured data only') and the same structured data enhanced with additional variables derived from unstructured documents ('structured and unstructured data'). Patient and disease features characterized included gender, age at advanced diagnosis, stage at diagnosis, length of follow-up, histology, smoking status, EGFR tested, EGFR status among those tested, *ALK* tested and *ALK* status among those tested.

Derivation of the overall study population is outlined in Table 1. There were 26,630 people in the EHR generated database with an ICD-9 code signaling lung cancer and two visits on or after 2013 (data through 31 August 2015). When only the structured data are used to select people with metastatic NSCLC, a total of 3562 people were in the final cohort; identifying patients with unresectable but otherwise nonmetastatic disease (e.g., stage IIIB) is not possible, nor is identifying the lung cancer histology. When the unstructured data are combed to confirm relevant characteristics, a total of 8324 people with advanced NSCLC are identified including confirmation of NSCLC with histology and inclusion of people with IIIB, IV and recurrent disease. The true cohort of interest – patients with advanced squamous NSCLC – was only assessable in the structured and unstructured data cohort since identification of these patients requires review of unstructured pathology documents to confirm histology. Of the 8324 people with advanced NSCLC, 2092 have advanced squamous NSCLC.

Among the 8324 people in the cohort generated using structured and unstructured data, only 2472 were also in the cohort generated using structured data only. Furthermore, 1090 people were included in the Structured data only cohort that are unlikely to meet the true parameters of the study population and would be erroneously included in an analysis that only uses structured data to select the study population (Figure 1). As patients were abstracted over a period of time, it is possible that some of the patients in the structured data only cohort subsequently developed advanced disease, thus potentially overestimating this false-positive rate.

Population characteristics are summarized in Table 2. Only patients for whom unstructured data was processed are described in Table 2 to ensure a fair comparison of patient

Table 1. Comparison of cohorts generated using structured electronic health record data only versus structured electronic health record data supplemented with abstracted unstructured data.

Goal	Structured data only	Structured and unstructured data
Recent LC patients	ICD-9 code of 162.x with at least two visits ≥2013 (n = 26,630)	ICD-9 code of 162.x with at least two visits ≥2013 (n = 26,630)
NSCLC patients	Patients without an administration for etoposide (n = 23,235)	Patients with confirmed NSCLC (n = 21,445)
Advanced NSCLC patients	Patients with a diagnosis for secondary metastases (ICD9 196.x–198.x) (n = 4382)	Patients with a confirmed diagnosis of advanced NSCLC (n = 10,826)
Patients with an advanced diagnosis date after 2013	Patients with a first diagnosis for secondary metastases ≥2013 (n = 3562)	Patients with a confirmed date of advanced diagnosis ≥2013 (n = 8324)
Squamous cell NSCLC patients	Unable to distinguish	Patients with a confirmed diagnosis of squamous cell carcinoma (n = 2092)

LC: Lung cancer; NSCLC: Non-small-cell lung cancer.

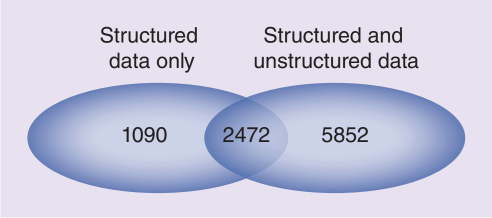


Figure 1. Comparison of patients selected for the analysis using structured data only versus structured and unstructured data.

characteristics. Therefore, **Table 2** shows the patient characteristics for all patients who could only be identified using structured and unstructured data [3,8,24] and the subgroup of these patients who would have also been identified if using structured data only (n = 2472). As shown in **Table 2**, using the unstructured data allowed for the identification of more early stage patients who went on to develop advanced disease as well as more squamous cell carcinoma cases.

The primary outcome of interest was median survival time. Date of death was generated using Flatiron Health's proprietary mortality dataset which combines internal data and externally linked data sources. When looking at the structured data cohort (n = 3449), the median overall survival (OS) is 0.74 (95% CI: 0.70–0.80) years; this compares to a median OS of 0.99 (95% CI: 0.96–1.03) years for the structured and unstructured data cohort (n = 8235) **(Figure 2)**. The OS results in the structured and unstructured data cohort are consistent with estimates for OS in an NSCLC as documented in the published peer-reviewed literature [25]. This makes good sense because people with small cell lung cancer have shorter survival and the structured data cohort only focuses on people with an ICD9 code for metastatic disease. Meanwhile, the structured and unstructured data cohort eliminates small-cell lung cancer and includes the relevant group of people with advanced disease inclusive of earlier stages such as stage IIIB. Furthermore, the OS estimate for the structured and unstructured data cohort is in line with the literature [24]. The structured and unstructured data population is clearly the more relevant group for subsequent analyses.

• Case study: assessment of biomarker testing patterns & outcomes in an NSCLC patient population

Evidence-based guidelines recommend molecular testing for appropriate patients with advanced NSCLC at the time of diagnosis. However, patterns for biomarker test ordering and implementation can be complex and variable across regions, centers and health system types. Much is yet unknown about patient outcomes as it relates to timing of molecular testing [25]. We conducted a real world study to evaluate the patient demographic and clinical characteristics, healthcare resource utilization and survival outcomes associated with early molecular diagnostic testing (EMT) compared with delayed molecular testing (DMT) among patients with metastatic NSCLC treated in a community setting.

This study used clinical data from US Oncology's iKnowMed (iKM) oncology-specific EMR system. This system captures demographic, clinical and treatment data for patients receiving care within the US Oncology's network of approximately 1200 community-based oncologists. During the study time period, the iKM EMR system was implemented across approximately 82% of the US Oncology network.

While the primary data elements were extracted from the structured fields in the EHR, in order to identify the cohort of interest, a significant number of clinical variables had to be obtained from 'unstructured' fields through manual extraction, in other words, a modified chart review. Examples of these variables included smoking status, line of therapy, histology, molecular testing orders and results, and timing of testing and treatments. Once this dataset was compiled, the patients' clinical data was matched with their records of resource utilization from a proprietary claims data warehouse (CDW), thus providing an enriched dataset in which we could evaluate clinical and economic outcomes. The CDW repository consists of claims for services within the US Oncology network and includes common procedure coding system and common procedural terminology codes, date of service, quantity, amounts billed and primary payer. Pharmacy data from Care Advantage Specialized Pharmacy and on-site pharmacies were leveraged in this analysis. Further, the Social Security Death Index was used to assess the vital status (death) of patients. The social security number was used to link patients in both the oncology EHR database and the Social Security Death Master File developed by the Social Security Administration. The linkage was done by non-study personnel in compliance with HIPAA.

Of the 11,615 patients with NSCLC identified in the EHR database, 350 met the inclusion/exclusion criteria and a 2:1 case–control method was utilized, resulting in 249 (71%) EMT patients and 101 (29%) DMT patients **(Figure 3)**. The largest source of attrition was presence of molecular testing documented in EHR and 6 months follow-up.

The mean age of the total patient population was 64 years and 56% were female. Most of the patients had stage IIIB or stage IV disease. In terms of completeness of other patient characteristic information obtained through chart review, smoking status (with 70% history of), histology (93% adenocarcinoma) and payer type (48% Medicare, using chart review and CDW) were available in over 95% of patients. Using the structured data only, TNM staging was missing in 23% of patients and tumor grade was available for 65% of patients. Presence of a positive ALK gene rearrangement or *EGFR* mutation was identified in 112 (32% of tested sample) patients; 81 EMT and 31 DMT. Of the 249 patients in the EMT group, EGFR molecular test results showed that 64% were *EGFR*

Table 2. Baseline characteristics – patients that would have been uniquely identified only when using the combination of structured and unstructured data (n = 5852) versus patients who would be identified using both methods, structured data only and the combination of structured and unstructured data (n = 2472).

Baseline characteristic	Total	Patients identified using only unstructured data (n = 5852)	Patients identified using both methods (n = 2472)	p-value[†]
Age at advanced diagnosis (years):				
– Median (IQR)	69.0 (61.0–77.0)	70.0 (62.0–78.0)	67.0 (60.0–75.0)	<0.0001
Follow-up from advanced diagnosis (days):				
– Median (IQR)	162.0 (63.0–343.2)	154.0 (58.0–335.0)	181.0 (78.0–358.0)	<0.0001
Gender, n (%):				0.16
– Female	3983 (47.8)	2830 (48.4)	1153 (46.6)	
– Male	4341 (52.2)	3022 (51.6)	1319 (53.4)	
Smoking status, n (%):				0.20
– History of smoking	7046 (84.6)	4976 (85.0)	2070 (83.7)	
– No history of smoking	1025 (12.3)	709 (12.1)	316 (12.8)	
– Unknown/not documented	253 (3.0)	167 (2.9)	86 (3.5)	
Histology, n (%):				<0.0001
– Non-squamous cell carcinoma	5723 (68.8)	3872 (66.2)	1851 (74.9)	
– Squamous cell carcinoma	2092 (25.1)	1616 (27.6)	476 (19.3)	
– NSCLC histology NOS	509 (6.1)	364 (6.2)	145 (5.9)	
Group stage, n (%):				<0.0001
– Group stage in not reported	351 (4.2)	279 (4.8)	72 (2.9)	
– Stage I	594 (7.1)	469 (8.0)	125 (5.1)	
– Stage II	352 (4.2)	276 (4.7)	76 (3.1)	
– Stage III	1578 (19.0)	1356 (23.2)	222 (9.0)	
– Stage IV	5449 (65.5)	3472 (59.3)	1977 (80.0)	
EGFR status, n (%):				<0.0001
– Mutation negative (wild-type)	3533 (42.4)	2360 (40.3)	1173 (47.5)	
– Mutation positive	658 (7.9)	410 (7.0)	248 (10.0)	
– Not tested	3852 (46.3)	2888 (49.4)	964 (39.0)	
– Pending/unknown	110 (1.3)	87 (1.5)	23 (0.9)	
– Unsuccessful/indeterminate test	171 (2.1)	107 (1.8)	64 (2.6)	
ALK status, n (%):				<0.0001
– *ALK* negative/not detected	3611 (43.4)	2373 (40.6)	1238 (50.1)	
– *ALK* positive	135 (1.6)	90 (1.5)	45 (1.8)	
– Not tested	4184 (50.3)	3123 (53.4)	1061 (42.9)	
– Pending/unknown	139 (1.7)	103 (1.8)	36 (1.5)	
– Unsuccessful/indeterminate test	255 (3.1)	163 (2.8)	92 (3.7)	

[†] For continuous variables, p-values correspond to t-tests between groups for normally distributed variables and Wilcoxon tests for all other variables. For categorical variables, p-values correspond to Chi-squared tests between groups or Fisher's exact test in the case of small cell sizes.

IQR: Interquartile range; NOS: Not otherwise specified; NSCLC: Non-small-cell lung cancer.

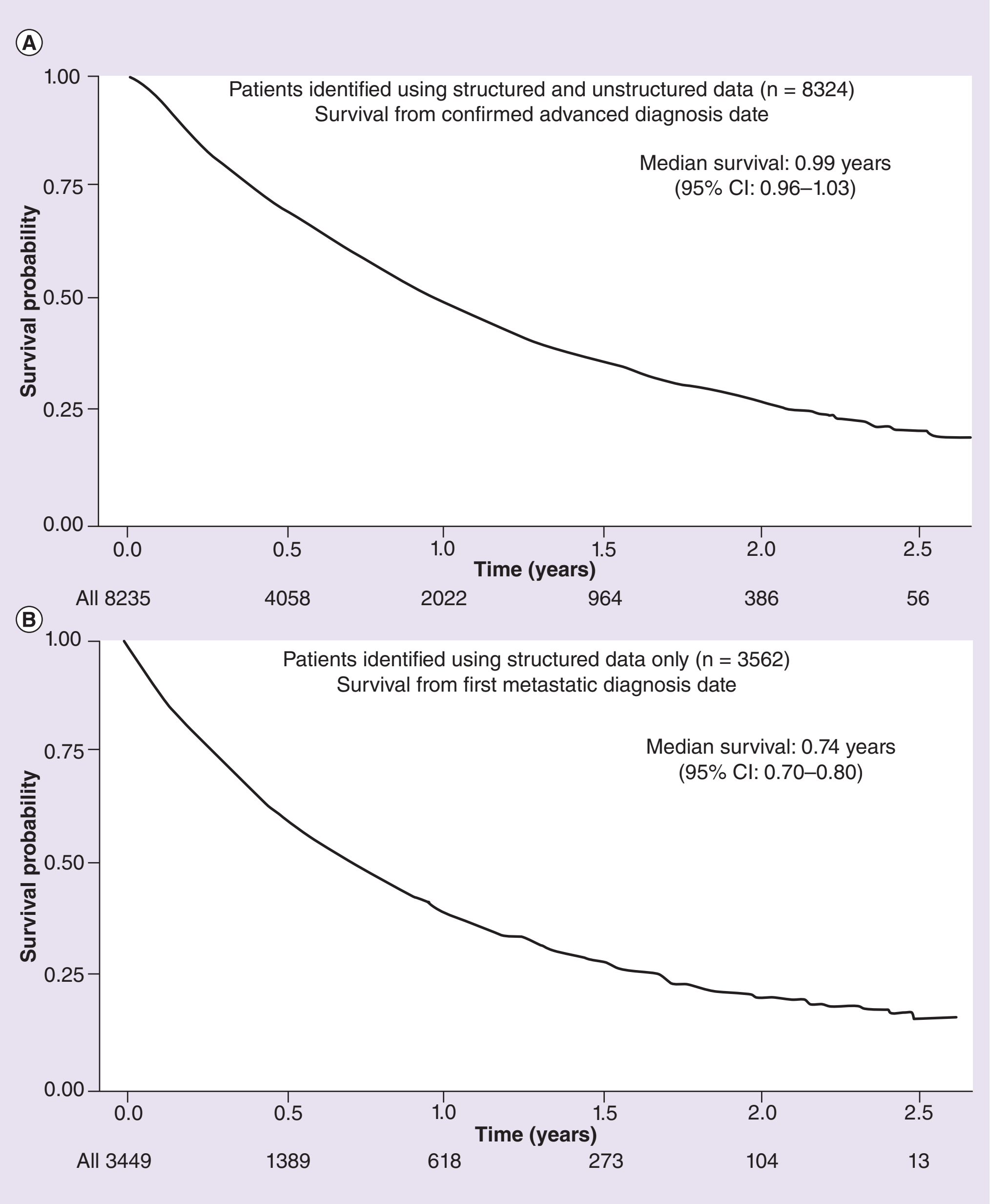


Figure 2. Comparison of survival curves using structured data only versus structured and unstructured data. Structured data are coded. Unstructured data are recorded as free text. Death date was generalized to the first of the month; In the top panel, 89 patients were removed from this analysis due to inconsistencies in dates (advanced diagnosis after death month). In the bottom panel, 113 patients were removed from this analysis due to date inconsistencies where the assumed date of metastatic disease was after the month of death.

wild-type (negative), while others were either positive (28%) or results were unknown due to insufficient sample (2%) and other reasons [6%]. The ALK molecular test results showed that 60% were ALK wild-type (negative) and others were either positive (5%), or results were unknown due to insufficient sample (2%) and other reasons (33%).

In terms of outcomes, the respective median OS (19.7 vs 22.3 months) and PFS (18.0 vs 18.1 months) for EMT and DMT were comparable between cohorts. Interestingly, across all multivariate Cox Proportional Hazards Regression models for OS and PFS, performance status was found to be independently associated with the risk of time to death and time to progression or death. Further, patients undergoing EMT were nearly three times more likely to exhibit poorer performance status compared with those undergoing DMT. Yet, performance status was missing or unknown in 19% of the overall sample and up to 24% in the DMT.

Despite multiple biopsies being more common in the cohort of patients receiving EMT, in the 273 (78%) patients with linked CDW data, cohorts were not significantly different in mean health resource utilization including inpatient and outpatient visits, chemotherapy and nonchemotherapy drug use, G-code, nursing home/hospice, laboratory, minor procedures, radiology and other services. Interestingly, 57% of patients were recorded as not receiving radiation therapy in the medical record, but the average radiation treatment was 1.4–1.7 and radiation service was 1.2–1.3 per patient per month. In a multivariate Poisson regression analysis, there was some evidence of an association between performance status and outpatient visits and inpatient services while stage and gender were significantly associated with radiation treatment.

In conducting this study, we noted significant gaps in terms of recording of molecular testing

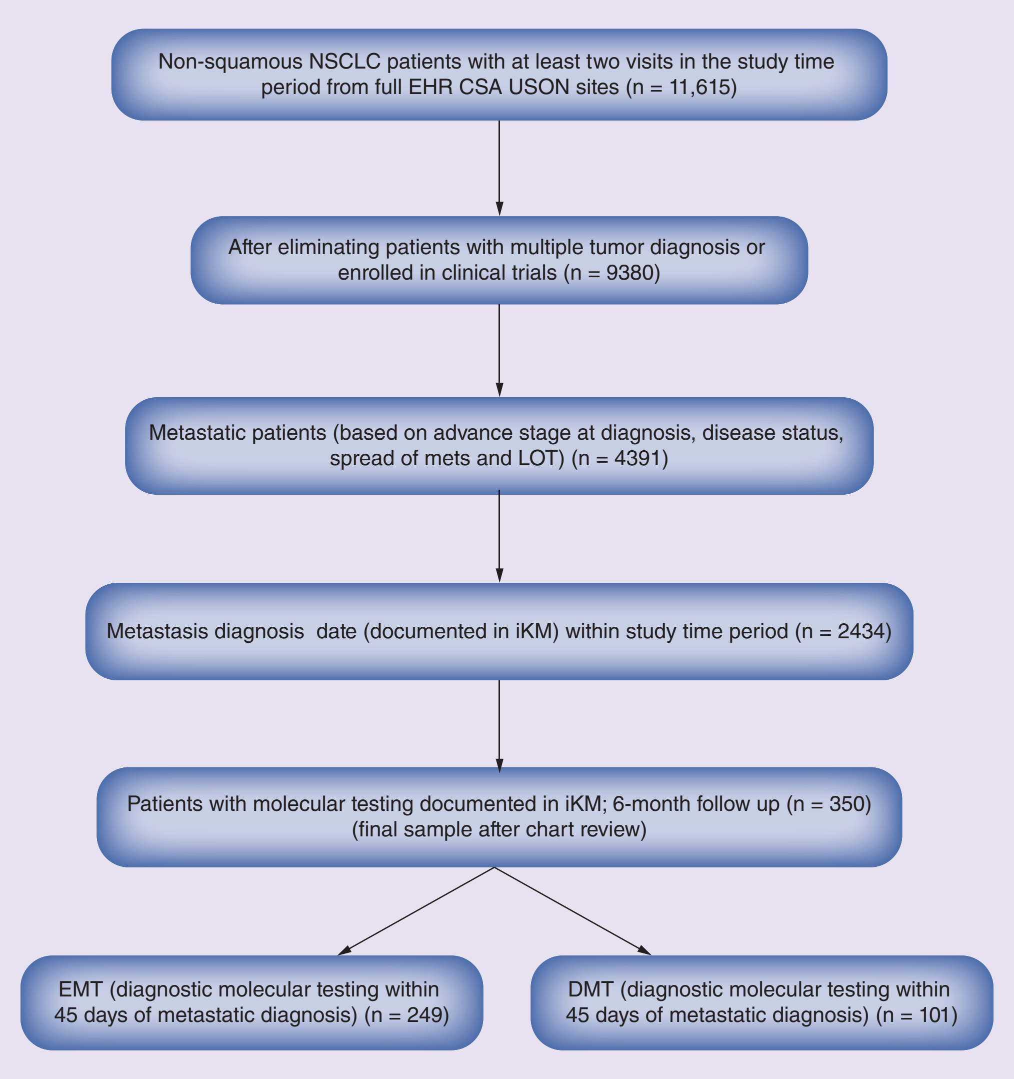

Non-squamous NSCLC patients with at least two visits in the study time period from full EHR CSA USON sites (n = 11,615)
After eliminating patients with multiple tumor diagnosis or enrolled in clinical trials (n = 9380)
Metastatic patients (based on advance stage at diagnosis, disease status, spread of mets and LOT) (n = 4391)
Metastasis diagnosis date (documented in iKM) within study time period (n = 2434)
Patients with molecular testing documented in iKM; 6-month follow up (n = 350) (final sample after chart review)
EMT (diagnostic molecular testing within 45 days of metastatic diagnosis) (n = 249)
DMT (diagnostic molecular testing within 45 days of metastatic diagnosis) (n = 101)

Figure 3. Patient selection procedure.
CSA: Comprehensive Strategic Alliances (practice management affiliation with USON); DMT: Delayed molecular testing; EHR: Electronic health record; EMT: Early molecular testing ; iKM: iKnowMed (USON's proprietary EHR system); LOT: Line of Therapy; NSCLC: Non-small-cell lung cancer; USON: US Oncology Network.

data, resulting in a substantial exclusion of subjects from the study; thus our sample size was limited. Most information related to molecular testing is still being managed via paper and faxing and only entered into the patient's chart as free text notes or stand-alone PDF files. Additionally, we observed issues with completeness of data in EHRs, even in structured fields as previously described, and there was significant variability between practices and individual physicians in terms of recording basic clinical characteristics. As noted in the first case, this necessitates additional time and cost to compile the variables necessary to conduct observational studies. These gaps in molecular testing results and clinical characteristics such as performance status are a significant limitation since they are associated with survival and healthcare resource utilization outcomes. Linking of the structured and unstructured data variables in the EHR to additional data sources including the Social Security Death Master File and CDW were also critical to addressing the study objectives.

Conclusion & implications

EHRs hold great promise to revolutionize oncology clinical research and practice. As they exist today, however, our ability to effectively extract information from EHRs is limited. Work to harmonize structured data into a uniform analyzable format, while an important and significant effort, is not sufficient to extract the full value of data contained in the EHR. There is also a need for an integrated effort within the standards community to better express concepts in oncology. As illustrated in our case studies, use of unstructured data was imperative to accurately classify cases, treatment and evaluation of outcomes (e.g., OS). However, the current standard approach – chart review – does not allow unstructured data to be processed at scale. Newer approaches such as technology-enabled abstraction are moving us along the pathway toward scalable unstructured data processing, as demonstrated in the first case example. Application of an NLP to oncology EHRs has been reported and while promising is still early in its evaluation [26–28].

Improvements can be made to address the data challenge both in healthcare broadly and in oncology specifically. Data linkage between EHRs and additional datasets (e.g. next-generation sequencing, closed claims, among others), currently challenging under the constraints of The Health Insurance Portability and Accountability Act of 1996 (HIPAA), must be made easier. With improved data linkage, patient-level data will be much richer and can be analyzed to uncover trends that previously could not be examined. In both case studies, external data sources were used for capturing mortality in addition to the EHR. Our second case study also leveraged linked claims data and provided another dimension to the evaluation in terms of capturing healthcare resource use, which is important to payers and policymakers. To facilitate easier data linkage, there is the need to address the complexities surrounding de-identification. Under HIPAA, de-identification requirements are difficult to apply to large datasets. The 21st Century Cures Act aims to address this restriction by moving clinical research under the Treatment, Payment, and Health Care Operations exemption, which would allow for increased data sharing [29].

The advent of Big Data also means we need to be able to leverage it in near real time to optimize care in clinical practice. We must design rapid learning health systems, rather than simply stopping at the creation of new knowledge, that can transmit that knowledge to practitioners at the point of care. In addition, we must train those practitioners on how to use it. Duke University recently undertook such an initiative with its Learning Health System Training Program. This program aims to train resident physicians to integrate research findings into their clinical practice. Applied at a larger scale, this concept could drastically change the practice of evidence-based medicine [30].

In sum, we discussed the challenges of leveraging real world data that is captured by healthcare systems in EHRs and the importance of accessing unstructured data, especially for oncology, as well as data linkages for providing a more comprehensive characterization of patients, their treatments and outcomes. While patient privacy and data protection must always be a priority, EHR-derived data for research is critical to medical innovation. As the variety and size of these data sources and technologies to extract and analyze information continue to evolve, the ability to leverage these data and effectively apply novel learnings in tailoring treatments for individual patients offer promise in optimizing patient outcomes and reducing inefficiencies.

Future perspective

As new reimbursement models evolve with a focus on outcomes and affordability, or value,

EXECUTIVE SUMMARY

Background

- The promise of the Big Data era in healthcare has been tempered by challenges with both extracting data from electronic health records (EHRs) and linking that data to other valuable sources.
- This article focuses on these challenges and demonstrates the results that can be achieved when these problems are overcome.

General challenges of EHRs

- EHRs were not designed to facilitate health services research.
- Frequently, important clinical information is not documented or is not documented consistently. Additionally, information technology systems are highly fragmented, leading to information being stored in multiple disparate systems. Of importance, much of the critical clinical data in EHRs is unstructured (i.e., entered as free text or in scanned documents), making it difficult to extract at scale.
- These problems are augmented in oncology where many of the crucial variables to oncology research (e.g., biomarkers, disease staging) are recorded in unstructured data and/or require clinical input to make sense of them.

Achieving the vision of Big Data for oncology

- As larger digitized datasets become available, the promise of Big Data in oncology becomes more possible. Genomic, claims, patient-reported and other data can be linked to the EHR to drive outcomes-focused analyses.

Practical applications: cohort selection is the most critical challenge in analyzing EHR data

- In analyzing EHR data, cohort selection is the most fundamental task, because inappropriate cohorts lead to inaccurate results.
- The case studies outlined in this article use two examples from non-small-cell lung cancer (NSCLC) to demonstrate how combining structured plus unstructured EHR data, and linking EHRs to other data sources, can lead to meaningful improvement in these types of analyses.

Case study: assessing the management of NSCLC in the community with Flatiron technology

- This case study compares the use of structured data only versus structured and unstructured data to identify patients in a nationally representative EHR-based dataset with advanced squamous NSCLC.
- Using the combination of structured and unstructured data, 8324 patients were identified from the broad cohort as having advanced NSCLC. Of the 8324 patients, only 2472 were also in the cohort generated using structured data only. Furthermore, 1090 patients would be included in the structured data only cohort who should have been excluded based on additional unstructured data elements.
- This study demonstrates the importance of being able to access both structured and unstructured data sources to correctly select an analytic cohort.

Case study: assessment of biomarker testing patterns & outcomes in an NSCLC patient population

- This case study evaluated patient, demographic and clinical characteristics, healthcare resource utilization, and survival outcomes associated with early molecular diagnostic testing compared with delayed molecular testing amongst patients with metastatic NSCLC. Structured and unstructured EHR data, which were used to identify 350 (249 early molecular diagnostic testing, 101 delayed molecular testing) patients meeting the eligibility criteria, were linked with resource utilization data from a claims data warehouse and with death data from the Social Security Death Index.
- Cohorts were comparable in overall survival, progression-free survival and healthcare resource utilization.
- The case study demonstrates the need to link EHR data with other sources to tie clinical and demographic characteristics to resource utilization and outcomes.

Conclusion & implications

- EHRs hold great promise to revolutionize oncology clinical research and practice. As they exist today, however, our ability to effectively extract information from EHRs is limited.

10.2217/fon-2015-0043

EXECUTIVE SUMMARY (CONT.)

Conclusion & implications (cont.)

- As the variety and size of data sources and technologies to extract and analyze information continue to evolve, the ability to leverage these data and effectively apply novel learnings by tailoring treatments to individual patients offer promise in optimizing patient outcomes and reducing inefficiencies.

Future perspective

- As new reimbursement models evolve with a focus on outcomes and affordability, meaningful use of EHRs will become increasingly important for many purposes. As one example, payers and manufacturers are currently evaluating novel pricing and reimbursement models based on real-world evidence. Success of these arrangements will require accessible, analysis-ready patient-level linked data incorporating EHR and other datasets available in real time.

meaningful use of EHRs will become increasingly important and ideally provide more complete data. Numerous examples are emerging. For example, there is much discussion about the linkage of physician reimbursement to treatment outcomes for patients with cancer. In the USA, one of the largest health insurers, United Healthcare, conducted a 3-year pilot study evaluating implementation of a new cancer care payment model rewarding physicians for recommended treatment practices and health outcomes which resulted in a 34% reduction in medical costs while sustaining quality of care [31]. The Center for Medicare and Medicaid Innovation Center (CMMI) is developing an Oncology Care Model for financial incentives, including performance-based payments, to improve care coordination, appropriateness of care and access for beneficiaries undergoing chemotherapy [32]. In October 2015, the National Comprehensive Cancer Network included drug affordability as one of five measures (efficacy, safety, quantity/quality and consistency of evidence) in its clinical practice guidelines that initially focus on multiple myeloma and chronic myeloid leukemia [33].

In parallel, payers and manufacturers are evaluating novel pricing and reimbursement models. For example, the largest pharmacy benefit manager in the USA, Express Scripts Inc., announced their interest in partnering with manufacturers on an indication specific pricing model that would provide differentiated oncology drug pricing based on efficacy across different tumor types [34]. Outside the USA, performance-based schemes have been introduced in oncology. In the UK, the National Health Services worked with the manufacturer of Velcade (bortezomib); as part of the agreement, the insurer would be reimbursed for the first four cycles of treatment if there is no patient response (≥25% reduction in serum M protein) [35]. In Italy, Jarosławski *et al.* identified 12 oncology risk sharing agreements out of 19 overall [36]. The Italian Medicines Agency's (AIFA) Oncologic Working Group created two types of risk sharing schemes that yielded discounts (50% from current prices for an agreed number of cycles) and/or refunds for patients who did not respond to therapy [37].

At the heart of these arrangements is the need for clinical information to identify patients of interest and clinical outcomes to execute and manage the agreements. All of these examples will require accessible, analysis-ready patient-level linked data incorporating EHR and other datasets available in real time. In fact, few agreements have been reported, especially in the USA, because of the complexity of implementing these contracts, especially with regard to availability of required information [35]. Even for the risk-sharing agreements in Italy that leveraged registries, Jarosławski *et al.* noted reports that only half of eligible patients in some regions were captured in those registries [36]. In the USA, payer claims data is often used for evaluating treatments and outcomes such as hospitalizations. For oncology, use of payer claims data alone and even with structured EHR is not sufficient. Thus, implementation of novel reimbursement models linked to outcomes, while promising, will need to address health information technology infrastructure including integration and interoperability across data sources and settings, incentives for physicians to document outcomes and critical information and the ability to rapidly access the derived data by participating parties for evaluation purposes. Solving the Big Data equation in oncology is paramount for these value-based payment models.

Financial & competing interests disclosure

M Berger, G Smith and J Harnett are employees of Pfizer, Inc. Pfizer, Inc. sponsored the research cited in the second case study. M Curtis and A Abernethy are employees of

Flatiron Health. Flatiron Health sponsored the research cited in the first case study. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

1 Howie L, Hirsch B, Locklear T, Abernethy AP. Assessing the value of patient-generated data to comparative effectiveness research. *Health Aff.* 33(7), 1220–1228 (2014).

2 Vassar M, Holzmann M. The retrospective chart review: important methodological considerations. *J. Educ. Eval. Health Prof.* 10(12), 1–7 (2013).

3 Manasanch EE, Smith JK, Bodnari A *et al.* Tumor registry versus physician medical record review: a direct comparison of patients with pancreatic neuroendocrine tumors. *J. Oncol. Pract.* 7(2), 111–116 (2011).

4 Clegg LX, Feuer EJ, Midthune DN, Fay MP, Hankey BF. Impact of reporting delay and reporting error on cancer incidence rates and trends. *J. Natl Cancer Inst.* 94(20), 1537–1545 (2002).

5 Rosenbaum L. Perspective: transitional chaos or enduring harm? The HER and the disruption of medicine. *N. Engl. J. Med.* 373(17), 1585–1588 (2015).

6 Brookstone A. HIMSS13 – R satisfaction diminishing. www.americanehr.com

7 Koopman RJ, Steege LM, Moore JL *et al.* Physician information needs and electronic records (EHRs): time to reengineer the clinic note. *J. Am. Board Fam. Med.* 28(3), 316–323 (2015).

8 Fernandopulle R, Patel N. How the electronic health record did not measure up to the demands of our medical home practice. *Health Aff.* 29(4), 622–628 (2010).

9 LeBlanc TW, Back AL, Danis M, Abernethy AP. Electronic health records (EHRs) in the oncology clinic: how clinician interaction with EHRs can improve communication with the patient. *J. Oncol. Pract.* 10(5), 317–321 (2014).

10 *Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary.* Grossman C, Powers B, McGinnis JM (Eds). National Academies Press, Washington, DC, USA (2011).

11 Vest JR, Gamm LD. Health information exchange: persistent challenges and new strategies. *J. Am. Med. Inform. Assoc.* 17, 288–294 (2010).

12 American Society of Clinical Oncology. Cancer-specific data sharing standards for communication, collaboration, and coordination of care. www.asco.org

13 Kite BJ, Tangasi W, Kelley M, Bower JK, Foraker RE. Electronic medical records and their use in health promotion and population research of cardiovascular disease. *Curr. Cardiovasc. Risk Rep.* 9, 422–425 (2015).

14 Kohane IS. Using electronic health records to drive discovery in disease genomics, *Nat. Rev. Genet.* 12, 417–428 (2011).

15 Walker AM, Zhou X, Ananthakrishnan AN *et al.* Computer-assisted expert case definition in electronic health records. *Int. J. Med. Inform.* 86, 62–70 (2016).

16 Carrell DS, Cronkite D, Palmer RE *et al.* Using natural language processing to identify problem usage of prescription opioids. *Int. J. Med. Inform.* 84(12), 1057–1064 (2015).

17 Vijayakrishnan R, Steinhubl SR, Ng K *et al.* Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. *J. Card. Fail.* 20(7), 459–464 (2014).

18 Byrd RJ, Steinhubl SR, Sun J, Ebadollahi S, Stewart WF. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *Int. J. Med. Inform.* 83(12), 983–992 (2014).

19 Zheng C, Rashid N, Koblick R, An J. Medication extraction from electronic clinical notes in an integrated health system: a study on aspirin use in patients with nonvalvular atrial fibrillation. *Clin. Ther.* 37(9), 2048–2058 (2015).

20 Raghavan P, Chen JL, Fosler-Lussier, Lai AM. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? *AMIA Jt Summits Transl. Sci. Proc.* 218–223 (2014).

21 Da Graca B, Filardo G, Nicewander D. Consequences for healthcare quality and research of the exclusion of records from the Death Master File. *Circ. Cardiovasc. Qual. Outcomes.* 6(1), 124–128 (2013).

22 Blackstone EH. Demise of a vital resource. *J. Thorac. Cardiovasc. Surg.* 143(1), 37–38 (2012).

23 Liede A, Hernandez RK, Roth M, Calkins G, Larrabee K, Nicacio L. Validation of international classification of diseases coding for bone metastases in electronic health records using technology-enabled abstraction. *Clin. Epidemiol.* 7, 441–448 (2015).

24 Ho C, Ramsden K, Zhai Y *et al.* Less toxic chemotherapy improves uptake of all lines of chemotherapy in advanced non-small-cell lung cancer: a 10 year retrospective population-based review. *J. Thorac. Oncol.* 9(8), 1180–1186 (2014).

25 Schink J, Trosman J, Weldon C *et al.* biomarker testing for breast, lung, and esophageal cancers at NCI designated cancer centers. *J. Natl Cancer Inst.* doi:10.1093/jnci/dju256 (2014) (Epub ahead of print).

26 Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J. Biomed. Inform.* 46, 830–836 (2013).

27 Warner JL, Levy MA, Neuss MN. Feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. *J. Oncol. Pract.* doi:10.1200/JOP.2015.004622 (2015) (Epub ahead of print).

28 Carrell DS, Halgrim S, Tran DT *et al.* Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *Am. J. Epidemiol.* 179(6), 749–758 (2014).

29 Robinson R. 21st Century Cures Act. Pharmavoice Nov/Dec 2015. www.pharmavoice.com

30 Abernethy AP, Etheredge LM, Ganz PA *et al.* Rapid-learning system for cancer care. *J. Clin. Oncol.* 28(27), 4268–4274 (2010).

31 Newcomer LN, Gould B, Page RD, Donelan SA, Perkins M. Changing physician incentives for affordable, quality cancer care: results of an episode payment model. *J. Oncol. Pract.* 10(5), 322–326 (2014).

32 CMS.gov. Center for Medicare and Medicaid Services. Oncology care model. http://innovation.cms.gov

33 Kelly C. NCCN Drug affordability ratings will start with multiple myeloma, leukemia. *The Pink Sheet* 7 September 2015. https://www.pharmamedtechbi.com

34 Loftus P. New Push Ties Cost of Drugs to How Well They Work. *The Wall Street Journal* 26 May 2015. www.wsj.com

35 Neumann PJ, Chambers JD, Simon F, Meckley LM. Risk-sharing arrangements that link payment for drugs to health outcomes are proving hard to implement. *Health Aff.* 30(12), 2329–2337 (2011).

36 Jarosławski S, Toumi M. Market access agreements for pharmaceuticals in Europe: diversity of approaches and underlying Concepts. *BMC Health Services Res.* 11, 259 (2011).

37 Adamski J, Godman B, Ofierska-Sujkowska G. Risk sharing arrangements for pharmaceuticals: potential considerations and recommendations for European payers. *BMC Health Services Res.* 10, 153 (2010).