Working with Oncology Electronic Medical Record (EMR) Data in Outcomes Research

Kathleen A. Foley, PhD, Senior Director, Global Health & Value Innovation Center, Pfizer; **Ronda Copher, PhD**, Associate Director, Eisai, Inc., Woodcliff Lake, NJ, USA; **Katherine B. Winfree, PhD**, Senior Research Scientist, Eli Lilly and Company, Indianapolis, IN, USA; and **Leigh G. Hansen, MS, MBA**, Vice President, Truven Health Analytics, Cambridge, MA, USA

KEY POINTS . . .

Electronic medical records (EMR) data have varying levels of completeness with regard to disease stage, biomarkers, adverse events, hospitalizations and emergency room visits, and survival.

While challenges exist in the use of EMR data for estimating survival outcomes, the benefits of such data are real and manageable with reasonable assumptions and analytical techniques.

When laboratory data are incorporated into EMR systems from electronic data feeds, they provide a consistent and reliable recording of one important component of a patient's clinical status.

Introduction

The use of clinical data from oncology-specific de-identified electronic medical records (EMR) is rapidly growing in the field of outcomes research. Cancer outcomes research and cancer data are distinct in that the data requirements for identifying a specific disease population are more stringent compared to other disease areas. In addition, specific clinical information is required to understand the disease state and treatment, which is not available in administrative claims data. As drug development increasingly focuses on personalized medicines, the need for data that include biomarkers and clinical details on disease state and progression are becoming more important for oncology outcomes research. These data are often found in EMRs.

The clinical evaluation of a patient with cancer offers multiple opportunities to capture information integral to the disease state and treatment. Patients with cancer typically undergo lengthy and complex diagnostic processes that often involve the use of cancer ICD-9 codes, which can occur even before a patient is completely diagnosed. Once diagnosed, disease stage and characteristics are critical factors in establishing an appropriate treatment plan, and these details are not captured in claims data. Progression and complications of the disease can be poorly recorded in standardized fields in both claims and EMR, making research in either data source a challenge. Oncology outcomes research using claims and EMR data is still in its infancy and the goal of this paper is to create a dialogue among interested researchers about important considerations and best practices when working with EMR data. Although oncology EMR data presents unique challenges in outcomes research, a deeper understanding of the data sources and limitations can help identify methodological approaches to handling these challenges.

Although oncology EMR data presents unique challenges in outcomes research, a deeper understanding of the data sources and limitations can help identify methodological approaches to handling these challenges.

Current Research Challenges Working with Oncology EMR Data

Despite the value of information obtained from EMR data, there are challenges. Oncology EMR data has limitations as information can be missing or incomplete, which can present methodological issues depending upon how the information gap is addressed. One common example of incomplete data capture or missing data occurs in patient histories; patients' previous treatments and comorbidities are often not fully recorded, which limits a complete understanding of what impacted current treatment decisions as well as treatment outcomes.

Another example involves biomarkers, which are valuable for segmenting patients with breast, colon, and lung cancers; however, this information is often lacking. While standard fields to capture biomarker status may exist in an EMR, they are typically not well populated by clinicians. Instead, physicians are more likely to record such information in the clinical notes, which due to HIPAA reasons, cannot be included within externally licensed extracts of an EMR. Similarly, adverse events (AEs) are also under reported, dependent on the condition and the setting of care in which the adverse event was treated. For example, when adverse events are treated in a hospital or emergency room, the treatment plan is unlikely to be recorded in the EMR and may be missed altogether. Even where AEs are treated by the primary oncologist in the clinic and recorded in the EMR, the AE severity or grade is not recorded in the standard fields. Some of this information may be captured in the clinician's notes; however, those are not included with the EMR data set and reconciliation via chart abstraction is a resource intense means to collect information that may not be available.

Data on patient survival and death are also limited, typically, because the inclusion of such information in licensable data is likely to violate HIPAA regulations for de-identification.

Costs are not a central component of EMR data; however, costs are often salient for understanding health outcomes. Within EMR systems, information on the cost of treatments is not typically present and when it is, refers to the billed amount, not the paid amount (pre-adjudicated billing data versus adjudicated claims data). One additional element that would be useful for understanding cancer is genomic detail. Genomic information is generally collected; however, it is not often included in standard EMR data fields. Knowing tumor mutations, e.g. BRCA1, BRCA2, p53, blood, and tissue genomics, would aid in more thorough understanding of which patients benefit most from which treatments, as well as help identify patients for clinical trials.

The Role of Identifying Observation Periods in Estimating Survival Endpoints

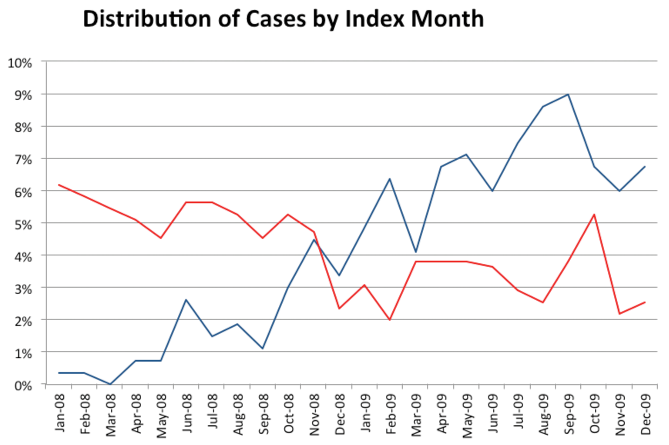
Overall survival (OS) remains the primary endpoint for most oncology clinical trials and for FDA approval. Therefore, it comes as no surprise that estimating survival using real-world data sources is of interest to many researchers. Drawing a link between OS in clinical trials and real-world settings is important to physicians, payers, and most importantly, patients. Since EMR data contains the clinical (i.e., stage of disease, performance status, histology, specific tumor type) and demographic (i.e., age, gender) characteristics required for accurate interpretation of survival data, it makes sense that researchers would look to use EMR data to estimate real-world survival.

When working with administrative claims data, the researcher has access to the patient's enrollment period within the data source – be it the employer or the health plan. Enrollment data provide identifiable start and stop dates for a patient's observation period, even if we do not know whether the reason for the end of a patient's observation is change of health plan, change of employer, or exit of employment due to disability or death. When working with EMR data, the beginning and end dates of observation are related to the patient's treatment at the specific clinic, which is related to disease state. With EMR data, the researcher cannot clearly identify a period prior to the disease, so as to isolate the beginning of the patients' disease periods, or follow the patients into hospital or hospice care, signifying the likely last phase of treatment prior to death. These limits on describing the observation period present a few methodological challenges for the researcher.

Given the lack of longitudinal detail for patients who come and go from a practice site based on the status of their disease and the frequent lack of date of death, one approach is to limit analyses to patients who have data for a minimum length of time. As with administrative data, however, this approach may introduce bias if sicker patients are systematically excluded from the analysis.

Censoring can affect the assessment of treatment outcomes in the real-world, especially when censoring differs by the treatment regimens. Comparing newly-approved treatments with older treatments means there is an inherent opportunity for longer follow-up among patients receiving older treatments. Alternatively, if guidelines change over the time period for which data are available, treatments under the new guidelines will have shorter follow-up periods than those under the previous guidelines. To determine whether differential follow-up due to month or year of treatment initiation is a potential problem, it is necessary to examine the index dates by treatment. Figure 1 illustrates an example graph of index dates. These data demonstrate the shift from one treatment to another over time, which would make a comparison of the follow-up or survival associated with the two treatments biased if index date was not taken into account in the analysis.

Figure 1: Index Data Exploration.



Differential censoring may also occur when longer survival with one treatment versus another is observed in the absence of confirmed dates of death. To limit the effect of these potential biases on cohort comparisons, consideration of methods is critical; matching by index date or extending the tails of Kaplan-Meier curves with extrapolation techniques may improve the accuracy of survival estimates, especially when considering mean survival. In Table 1, Cohort A has a greater percentage of censoring than the comparator cohorts. While the median and mean estimates are directionally similar, the higher censoring rate provides reason to further explore the data. In this case, an exponential extrapolation was applied to the mean of the observed data. The differences in mean survival among the cohorts become more apparent when extrapolating beyond the observed data to generate the extended mean.
Several other potential issues with EMR data, as with any real-world data source, also require careful consideration in any oncology EMR study. Treatment regimens may appear very different from the prescribed patterns tested in randomized controlled trials because clinical practice permits physicians to make treatment choices and trade-offs for their individual patients. In addition, lines of therapy may not be clearly differentiated or defined due to application of different algorithm rules by EMR data holders and lack of clearly defined dates of progression at which point

Table 1: Example of Observed and Extended Mean Survival in the Presence of Differential Censoring

	Cohort A	Cohort B	Cohort C	Cohort D
Observed median (months)	11	8	10	10
Observed mean (months)	13.9	12.5	13.2	13.1
Extended mean (months)	17.5	14.4	15.2	15
Delta mean (months)	3.7	1.9	1.9	1.9
% censored	33.4	19.9	21.2	20.4

>

Table 2: Example Patient Status Records and Chemotherapy Initiation

Patient 1	Patient 2	Patient 3	Patient 4
First Status: 2/8/2006 Relapse	First Status: 6/10/2005 Remission	First Status: 11/04/2011 Relapse	First Status: 08/30/2010 Relapse
Second Status: 11/18/2008 - Active	Second Status: 10/14/2008 - Relapse	No chemo	First Chemo: 09/10/2013 Rituximab
Third Status: 8/23/2011 Remission	Third Status: 08/25/2009 Partial Remission		
First Chemo 5/31/2013 Rituximab	First Chemo 1/19/2010 Bendamustine		

an escalation in line of therapy could be easily interpreted. For example, changing one drug within a three-drug regimen should not automatically trigger an escalation in line of therapy; a drug of the same class may be substituted for tolerability reasons while the treatment is still considered the same line of therapy in the mind of the physician and patient. As researchers working with clinical EMR data, we need to remember the original purpose of the data we are exploring. These lines of therapy are important considerations when attempting to estimate survival for patients at a certain point in their treatment plans, i.e., first line, second line.

While challenges exist in the use of EMR data for estimating survival outcomes, the benefits of such data are real and manageable with reasonable assumptions and analytical techniques.

Incomplete Disease Staging and Status

One of the key challenges with oncology EMR data is inconsistent recording in structured fields of key clinical characteristics such as disease stage, status, and progression, especially among patients with a hematologic cancer, for which many variations in these clinical descriptions exist relative to solid tumors. To accurately identify hematologic cancer patients for study inclusion where these important data are unavailable, algorithms for identifying patients who have relapsed or had refractory disease or disease progression are needed.

Progression status and disease response may not be well captured in structured variables within the EMR data. In the absence of progression status, progression may be estimated through an escalation in line of therapy. Some EMRs allow for electronic search of the progress notes in which “progression” may be mentioned; yet, progression status in the notes may not be consistently reported across patients given the different clinicians making use of the EMR system. Finally, a chart review may be possible for supplementing EMR data; however, this approach is often costly in terms of time and resources. Alternatively, it may be possible to develop algorithms using the data that does exist in the structured EMR variables.

There are two approaches for developing an algorithm: 1) work with a group of patients who have a reported clinical status, or 2) work with clinical data (i.e. patients’ symptoms, lab results, etc.). To explore these approaches, Truven Health Analytics undertook a pilot study using the Truven Health MarketScan® Oncology EMR Database for a sample of patients with chronic lymphocytic leukemia (CLL).

Approach 1: Patient Clinical Status

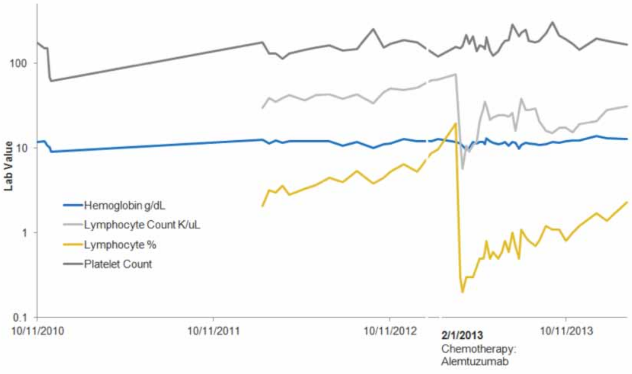
Most EMR systems have data fields to capture patient clinical status such as disease relapse, remission, refractory, chronic, active, and stable, among others. Of 18,334 patients identified with CLL in the EMR data, 7,865 (43%) had a patient status reported, and 528 had any mention of either relapse or remission. No records used the term ‘refractory’. Seventy-three (1%) patients had a record for relapse on the same date as a CLL diagnosis and no evidence of any other cancer types. The first level of exploration involved a review of all records for these 73 patients. Table 2 presents a summary of the recorded status for four patients from this cohort, as example records.

These patient examples demonstrate inconsistency between the patient status and the initiation of chemotherapy. In addition, the observed time between relapse date and chemotherapy initiation is significant, suggesting that patient status indicators are not updated regularly and may not necessarily be the driver of the decision to treat.

Approach 2: Patient Lab Records

Figure 2 presents an example of trends in laboratory values and initiation of chemotherapy for one patient. The trends demonstrate how initiation of chemotherapy is closely timed with increases in lymphocyte count and percent, along with a decrease in platelets and hemoglobin. The trends further demonstrate the rapid change in laboratory values following initiation of chemotherapy. Data for other patients showed similar patterns.

Figure 2: Lab Records Pre and Post Treatment (Single Patient).



2/1/2013
Chemotherapy:
Alemtuzumab

When laboratory data are incorporated into EMR systems from electronic data feeds, they provide a consistent and reliable recording of an important component of a patient’s clinical status. Following multiple lab results over time allows greater precision in identifying thresholds of patient changes that trigger treatment initiation. As illustrated above, this suggests that lab result changes noted just prior to chemotherapy initiation are consistent with guideline-based definitions of disease progression and relapse for CLL, and may provide a significant source of information for estimating patient clinical status as well as response to treatment [1].

Summary

Oncology EMR data present unique challenges for researchers due to the current lack of incentives for standardized reporting and use of discrete data fields by providers. Current challenges include:

- Under reporting of adverse event grade and treatment plans;
- Limited, if any, cost data;
- Under-reporting of patient history, clinical status, and disease progression;
- Incomplete capture of genomic information;
- Impact of censoring on estimates of overall survival and progression free survival; and
- Potential bias when continuous observation periods are used.

Challenges with censoring and observation windows can be addressed by matching on index dates and applying analytic techniques to account for differential censoring. Challenges with under-reporting of patient status and progression can potentially be addressed via algorithms based on laboratory results and use of ICD-9 codes for disease-related symptoms.

Oncology outcomes researchers are strongly encouraged to publish and share methodological work using oncology EMR data to facilitate the development and dissemination of best practices in this important area of research.

References

[1] Hallek M, Cheson B, Catovsky D, et al. Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute- Working group 1996 guidelines. Blood 2008;111:5446-56. ■

Additional information:
The preceding article was based on the workshop, "A Realistic Approach to Working with Oncology Electronic Medical Record (EMR) Data In Outcomes Research," presented at the ISPOR 19th Annual International Meeting, Montreal, QC, Canada, June 2, 2014.

To learn about the new ISPOR Oncology Special Interest Group, go to: http://www.ispor.org/sigs/Oncology.asp. Additional reading on electronic medical/health records can also be found in ISPOR's Reliability and Validity of Data Sources for Outcomes Research & Disease and Health Management Programs. Go to: http://www.ispor.org/publications/DataSorcesBook.asp for more information and for details on how to purchase copies of this valuable book.