Clinical Intervention Research with EHR
A Big Data Analytics Approach

Emergent Research Forum (ERF)

Rupesh Agrawal
Oklahoma State University
rupesh.agrawal@okstate.edu

Dursun Delen
Oklahoma State University
dursun.delen@okstate.edu

Bruce Benjamin
Oklahoma State University
bruce.benjamin@okstate.edu

Abstract

Electronic Health Record adoption has reached its saturation point, near 100%; however, it has not been fully utilized for clinical intervention research outcomes. Traditional clinical trial studies for medical intervention guidelines have proven to be time-demanding, expensive, with limited coverage and effectiveness. This study aims to discover conclusive evidence with high significance to the established clinical correlation between obstructive sleep apnea and related comorbidities, along with an application of analytics to establish the impact of obstructive sleep apnea on stroke risk rate using a large, feature-rich electronic health record database. Long term goal of this research is to use a variety of data sources, including electronic health record data, along with big data analytics tools to develop a new stroke-risk stratification score. The scoring system is expected to be able to generate the granular measures for a patient-centric dose-drug combination, consequently reducing the side effects while improving the treatment plan.

Keywords

Electronic Health Record, Data Quality, Clinical Trial, Analytics, OSA, AF

Introduction

In 2009, the American Reinvestment & Recovery Act, forced the adoption of Electronic Health Record (EHR) in hospitals. Within a decade adoption is close to 100%. However, EHR has not been harnessed to its full potential to enhance guidelines for clinical intervention and to conduct prospective clinical trials. The Goal of the Health Information Technology for Economic and Clinical Health Act of 2009 is to promote meaningful use of EHR for timely exchange of health information to improve the quality of care, safety, efficiency, and health disparities of patients. Many efforts have powered the promotion meaningful use. FDA has released a guide for use of EHR data for prospective clinical investigations to provide access to real-time data and clinical trial efficiency (FDA, 2018). Patient-centered outcome research institute (PCORI) has launched clinical data research networks to facilitate multi-site, observational, and interventional research (Fleurence et al., 2014). We argue that financial incentives and penalties had fueled the current adoption. For sustainable growth of EHR, we have to overcome the adverse perception of EHR quality, HIPPA concerns and improve the clinical outcome vis-à-vis better improved line of inquiry for the clinical trial, patient-targeted discovery for patient-centric evidence-based medicine. Data quality concern has demotivated researchers from ambitious interventional studies and to change the status-quo of clinical decision paradigm. Many groups have recognized the future impact of EHR, Consequently, use of EHR can bring the paradigm shift in clinical decision making with cost-effective, time-efficient, clinical research to deliver a unique solution (Raman et al., 2018).

Epidemiology-Prevalence Studies

The current clinical interventional studies, based on conclusions from epidemiological studies, measure the prevalence of the disease with a smaller representative sample of the target population. Per the
National Institute of Mental Health, “Prevalence is the proportion of a population who have a specific characteristic in a given period” (NIMH, 2017, Para#2). The conclusion from prevalence studies is susceptible to the error caused by estimation, instrument method, quality, and integrity of data (Andersson, 1999; Pearce, 2004; NIMH, 2017); independently epidemiological studies are rarely able to prove the cause of health symptoms (Farmer, R. 2007).

**Retrospective and Prospective Cohort Studies**

In the effort to gather the highest level of evidence to support the mission of evidence-based medicine, there is a hierarchy of studies. Each level of study increases the rigor, and the relevance to the study design by eliminating bias and error. A prospective study validates future health outcomes based on a line of inquiry or propositions, whereas retrospective study explores the past performance of health risk, disease symptoms. Prospective and Retrospective studies, collectively called as cohort studies, have inherent challenges of high cost, long study duration, weak result-significance, confounding bias, selection bias, recall bias (True Positive Rate), and low sample size (Barnett & Hyman, 2006; Hassan, 2006; Euser et al., 2009; StatsDirect, 2019).

**Electronic Health Record Data**

The data quality, integrity of the EHR data are current challenges to EHR-based interventional studies, along with concerns of various bias and data privacy challenges (Bowman, 2013; Hong et al., 2015; Weng, 2017; Raman et al., 2018). We argue that current level of quality and accuracy of EHR data is not at par, to support the full potential of the EHR system; however, there is constant awareness and general agreement to improve the current data quality expectation (Payne et al., 2015; Raman et al., 2018). For example, the data should be correct, and current (Zozus et al. 2014; Scholte et al., 2016; Weiskopf et al., 2017). Consequently, research-work, and industry initiatives have led the effort for improving data quality (Kahn et al., 2012; Dziadkowiec et al. 2014; Kahn et al., 2016; Qualls et al., 2018). Targeted initiatives a launched to tackle the data quality challenges, e.g., Observational Health Data Science and Initiative (OHDSI), funded by the National Institute of Health. OHDSI’s focuses on data standardization, data characterization, quality improvement, and PCORI, where the latter has developed standards for data integrity and rigorous analyses.

Subsequently, we propose that an EHR-based clinical trial study is a possibility, with educated inclusion-exclusion criteria from over four hundred clinical variables, from millions of records collected over a decade, and with a rich EHR data set representative of the US population. Using a new lens of patient-centric, evidence-based medicine, EHR data may resolve concerns for confounding bias. A patient has multiple health manifestations (diagnosed and undiagnosed), which are not captured, especially in high efficacy prospective-Randomized Control Studies (RCT) because of exclusion criteria to isolate the confounding errors. Consequently, Cohort and RCT studies have inherent selection bias by selecting healthiest patient data for obtaining high efficacy result in favor of the treatment under review (Raman et al., 2018), on the other hand, the data are subject to variations from different site-specific variation (Staman et al, 2017).

We argue to support the need for conducting an EHR based clinical trial study, followed by a traditional RCT study for validating the EHR study and this proposition is the foundation of our study goals. We collected de-identified, HIPAA-compliant electronic health records (EHR) from the Cerner HealthFacts® database for two different health conditions; 1) Obstructive Sleep Apnea (OSA), 2) Atrial Fibrillation (AF) to test following hypothesis:

1) H1: Prevalence estimates (comorbidities) are higher for epidemiological studies vs. EHR studies.

2) H2: Stroke risk rate of patient is higher with OSA comorbidity, than with just AF.

**Methodology**

Using structure query (SQL), 556, 474 unique patients diagnosed with OSA were extracted, from 9,118,897 clinical encounters of OSA patients, and 778, 746 unique patients diagnosed with AF were
extracted, from 12,028, 366 clinical encounters of AF patients. The patient records were extracted from HealthFacts® data repository of 63,382,356 unique patients. Encounters are the number of visits by all patients to a hospital or clinic facility. Patient comorbidities were identified per ICD9/10 code; OSA codes are 327.23/ G47.33, and AF codes are 427.3, I48, 427.31, I48.0, I48.1, I48.2, I48.9, I48.91. The extracted data were checked for accuracy, redundancy, missing values, dual coding and cleaned before use for final analysis.

For H1 Prevalence formula was used and numbers were plotted in Figure 2.

\[
\text{Prevalence} = \frac{\text{Number of people in sample with characteristic}}{\text{Total number of people in sample}}
\]

For H2, data was separated to represent groups (Figure 2). It also displays the size of each group. Factorial analysis is used for group mean difference. The results of ANOVA analysis is compiled in Table 1.

**Results**

Figure 2 and Table 1 show the preliminary results of the study. Figure 2 shows the comparison of prevalence numbers (%) between epidemiological studies and EHR study. Table 1 shows the results of 3-factor (Gender, Age, and Health Condition) ANOVA analysis of stroke risk score, and the difference in stroke risk score, showing an increase in score for patients with AF+OSA compared to a patient with AF.

**Table 1: ANOVA Analysis of Stroke Risk Score**

<table>
<thead>
<tr>
<th>Patient Characteristic</th>
<th>Risk Score</th>
<th>Difference</th>
<th>SS</th>
<th>NumDF</th>
<th>DenDF</th>
<th>F Ratio</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female, &lt;65, OSA</td>
<td>2.02</td>
<td>0.25</td>
<td>434.6</td>
<td>1</td>
<td>778734</td>
<td>1782.41</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Female, &lt;65, AF</td>
<td>1.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female, 65-74, OSA</td>
<td>3.05</td>
<td>0.15</td>
<td>153.1</td>
<td>1</td>
<td>778734</td>
<td>628.03</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Female, 65-74, AF</td>
<td>2.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female, &gt;75, OSA</td>
<td>4.1</td>
<td>0.08</td>
<td>51.3</td>
<td>1</td>
<td>778734</td>
<td>210.37</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Female, &gt;75, AF</td>
<td>4.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 1: Full Factorial ANOVA: Increase in Stroke Risk Score

<table>
<thead>
<tr>
<th>Group</th>
<th>β</th>
<th>SEβ</th>
<th>t-value</th>
<th>df</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male, &lt;65, OSA</td>
<td>0.99</td>
<td>0.22</td>
<td>681.8</td>
<td>1</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Male, &lt;65, AF</td>
<td>0.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male, 65–74, OSA</td>
<td>2.03</td>
<td>0.13</td>
<td>217.4</td>
<td>1</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Male, 65–74, AF</td>
<td>1.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male, &gt;75, OSA</td>
<td>3.07</td>
<td>0.09</td>
<td>111.1</td>
<td>1</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Male, &gt;75, AF</td>
<td>2.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion

The prevalence numbers (%) are significantly different (lower) using EHR dataset, compared to prevalence number (%) developed using traditional epidemiological studies (Figure 2). Currently, clinician use epidemiology-prevalence estimates to diagnose, and plan treatment. Consequently, the new lower prevalence numbers (%) sets the background for more in-depth analysis and may change the current status-quo of clinical diagnoses and treatment plan. Similarly, Table 1 shows that all patients groups with comorbidity (OSA) have a statistically significant and clinically higher stroke risk score. Traditional studies with smaller sample size can establish only significance, not clinical relevance. Consequently, we calculated the stroke risk score to show the increase in clinical effect size (relevance), i.e., the increase in the magnitude of the stroke risk score.

Conclusion

The current clinical decision making is based on an older paradigm, is time-consuming, based on smaller, and non-representative population data. Exponential growth in population representative volume rich data with variety, generated at high speed, has presented with new opportunities for clinical research. Time is apt for developing the body of knowledge using contemporary information science and technology methods, e.g., big data mining, and analytics, to solve patient-centric solutions for clinical intervention, that are cost-effective, efficient, relevant. One of the limitation of EHR based studies is inability to isolate patients with unreported comorbidities. EHR Data, although has few drawbacks (Perception, Data Quality) but can challenge the status-quo, to change the current paradigm of clinical intervention and decision making. We believe that under the vision of the current study a new refined scoring system for stroke-risk rate stratification will change the clinical decision-making paradigm. Besides, we can expand the IT artifact to study other health manifestation e.g., Type 2 Diabetes, COPD and cancer.

Acknowledgments

This work was conducted with the data from the Cerner Corporation’s HealthFacts data warehouse of electronic medical records, provided by Oklahoma State University (OSU), Center for Health Systems Innovation (CHSI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Cerner Corporation, OSU or CHSI.

REFERENCES

Hassan, E. (2006). Recall bias can be a threat to retrospective and prospective research designs. The Internet Journal of Epidemiology, 3(2), 339-412.