JOURNAL OF HEALTH CARE FINANCE

Spring 2017

Assessing Healthcare Utilization and Cost of Care Among Rheumatoid Arthritis Patients: A Comparison of Electronic Medical Records and Claims Data

Jing Hao, Ph.D
Assistant Professor
Department of Epidemiology and Health Services Research
Geisinger Health System
Danville, Pennsylvania USA

Daniel D. Maeng, Ph.D (Corrsponding Author)
Assistant Professor
Department of Epidemiology and Health Services Research
Geisinger Health System
Danville, Pennsylvania USA

Xiaowei Yan, PhD
Statistician Investigator
Sutter Health
Walnut Creek, California USA

Walter F. Stewart, Ph.D, MPH
Chief Research Officer
Sutter Health
Walnut Creek, California USA

Joseph A. Boscarino, PhD, MPH
Professor
Department of Epidemiology and Health Services Research
Geisinger Health System
Danville, Pennsylvania USA

Abstract

Electronic medical records (EMR) and health insurance claims data offer two potential data sources for researchers to examine healthcare utilization patterns and the cost of care. In particular, combining the clinical and epidemiological variables typically available in EMR with cost information available in the claims data is not only intuitively sensible, but also increasingly more feasible with growing standardization of EMR across healthcare delivery systems. In this study, we compare EMR and claims data within a cohort of rheumatoid arthritis patients who received care from Geisinger Health System (GHS) and also had concurrent Geisinger Health Plan (GHP) coverage. We also develop a cost "imputation" method to obtain GHP claims-based cost estimates within EMR, even for those who did not have GHP coverage. The findings confirm that there is significant disagreement between EMR and claims data and suggest that each represent a different set of clinical phenomena. This study also illustrates different factors to consider for researchers in choosing one data source over the other in conducting clinical research.

Introduction

In analyzing healthcare utilization and cost of care, two most frequently used data sources are health insurance claims and electronic medical records (EMR). In theory, both data sources can potentially yield the same information and allow the researcher to reach the same conclusion. For example, if a patient visits a primary care office and receives a series of diagnostic tests, each of these tests should be reflected in the EMR, and to the extent that the physician's office was reimbursed for all the services rendered, the patient's health insurance claims data should also reflect these same transactions.

In reality, however, it has been well-established that there are consistent and significant disagreements between what EMR capture versus what the corresponding claims data capture.[1-4] As such, relative usefulness of either of these data sources depends on the type of care under consideration and the specific research questions to be answered by the researcher.[4] Furthermore, with heightened focus on cost of care in recent years, there is increasing interest in combining clinical and epidemiological variables (e.g., comorbidity, disease severity/activity, biometric measurements such as body-mass index, blood pressure, hemoglobin A1c, etc.), as well as socio-economic variables (e.g., work status, race/ethnicity, and income) with economic variables, such as cost of care. Using EMR to analyze patterns of healthcare utilization and cost that reflect clinical and biological factors of each individual patient is not only intuitively sensible but is becoming more feasible as standardization of EMR across large healthcare delivery systems become more common.[5,6]

Although appealing, using EMR for study of utilization and cost is challenging in practice. While EMR have detailed clinical information on each patient, they lack cost information. Charge amounts that are sometimes available in EMR do not represent the actual payments received by the provider, which is typically substantially lower than the charge amounts.[5] Claims, on the other hand, have detailed cost data, but they have, at best, incomplete clinical information about the patient.[6] One solution is to focus on the subset of the patient population of interest who appear in both EMR and claims data. However, augmenting one data source – or filling in the blanks, so to speak – using another typically involves laborious manual review of individual patient records, and this approach is usually not feasible for studies that involve hundreds or thousands of patients. Furthermore, focusing only on this overlap between EMR and claims is likely to substantially reduce the sample size, leading to lower statistical power and generalizability of the analysis. As such, researchers are typically faced with using one data source over the other, usually without a means to determine whether their conclusions would be different had a different data source been used.

In this study, we compare EMR and claims data within a cohort of rheumatoid arthritis (RA) patients who received care from Geisinger Health System (GHS) between 1999 and 2011 *and* had concurrent Geisinger Health Plan (GHP) coverage. Because this cohort allows us to examine both EMR and claims data of the same individuals, we can examine whether claims and EMR produce the same estimates and, if not, to what extent they disagree with each other. We also develop a cost "imputation" method to obtain GHP claims-based cost estimates within EMR, even for those who did not have GHP coverage. Thus, the findings in the current study will assess if there is significant disagreement between EMR and claims data and, if found, will suggest the potential reasons for this disagreement. We also discuss different factors to consider for health researchers in choosing one data source over the other in clinical research.

Most of the published literature on healthcare use and RA disease progression has focused on specialty care patients without a means to define a source population (thereby limiting the generalizability of the study findings) or has had limited access to longitudinal data that can be used to characterize disease progression. For example, previous work on disease progression after RA diagnosis included a model that stratified patients between self-limiting, persistent non-erosive, and persistent erosive arthritis at the time of the first patient visit.[7] Another previous study predicted erosiveness and the onset rate of new erosion for patients with early RA.[8] A more recent study tested a claims-based algorithm to serve as a proxy for the clinical effectiveness of RA medications over a 12 month period, among individuals for whom treatment with a new biologic agent or non-biologic disease-modifying RA agent was being initiated.[9]

Methods

Data

For the current study, we used a longitudinal dataset developed from GHS's electronic EMR data and claims data from GHP. The sample was limited to those patients who had both EMR and claims during the study period. In addition, we also excluded those patients who did not have a recorded encounter in the EMR during any 6-month period since their first encounter in the study period (i.e., non-continuous enrollment in GHS). This was done to minimize an obvious source of discrepancy between EMR and claims data: i.e., some patients might have missing data in EMR because they might have stopped receiving care from GHS and instead received care elsewhere (due to, for example, relocation, coverage changes, etc.). Our approach was consistent with other standard procedures in pharmacoepidemiology.[10] This resulted in the final sample size of 989 patients. The mean age of these patients was 66 years old, with interquartile range of 58-77. Seventy-three percent (73%) of these patients were female.

Cost Imputation Method

As noted above, EMR lacks cost information. To circumvent this problem, we have developed a regression-based cost imputation method based on claims data as outlined below:

1) First, we started by applying the same inclusion and exclusion criteria to both the claims and EMR data to select the eligible patient population;

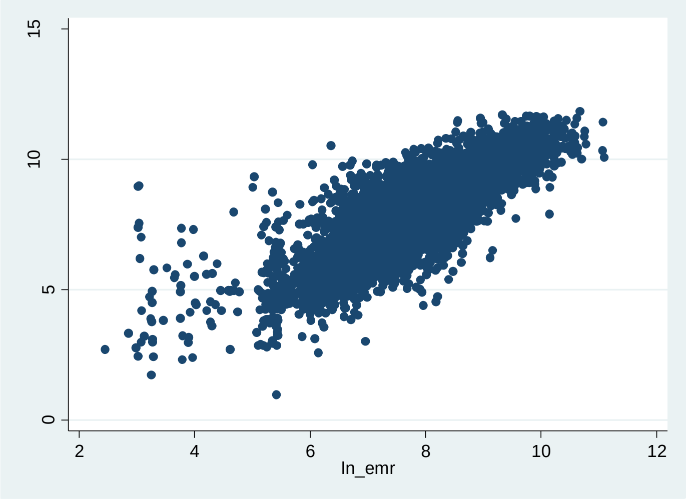
2) Second, we categorized encounter types in EMR and claims types into a set of mutually exclusive major categories. In this study, we used the following major categories: inpatient visit, outpatient visit, emergency department (ED), diagnostic imaging (i.e., X-rays, computerized tomography (CT), and magnetic resonance imaging (MRI)), and prescription drugs. Professional charges, which are typically available as separate claim types, were assumed to have been incurred in every encounter in EMR;
3) Third, in the claims data, we estimated the following multivariate regression model using Generalized Linear Model (GLM) with log link and gamma distribution function:
   Mean Cost = $\beta_0 + \beta_1$(Encounter Type) + β_2(Medicare) + β_3(Age) + β_4(Gender)
   "Encounter Type" denotes a set of binary indicator variables that represents each major encounter type category (e.g., inpatient, outpatient, ED, etc.); "Medicare" is a binary indicator variable that equals 1 if the patient has Medicare coverage and zero otherwise; "Age" is a continuous variable capturing the patient's age at the time of the study; and "Gender" captures the patient's gender.
4) Fourth, we took the beta coefficient estimates (obtained in step 3) and applied them to a similarly structured EMR data to obtain the estimated mean cost in the EMR.

The above method can be modified by introducing interaction effects between the encounter type variables and the age or the gender, for instance. In our estimates, the results are not sensitive to such alternative specifications. The resulting cost estimates can be interpreted as "imputed cost" under the hypothetical scenario that the patient had been covered by GHP. The same cost imputation method has been used previously elsewhere.[11,12]

The advantage of this cost imputation method is that it is not necessary that those patients who are included in the claims data be also included in the EMR data; as long as the structure of the EMR data can be modified to accommodate the above regression model, estimated cost can be obtained for that patient. The disadvantage of this method is that its accuracy may depend on the potentially subjective categorization of claim and encounter types.

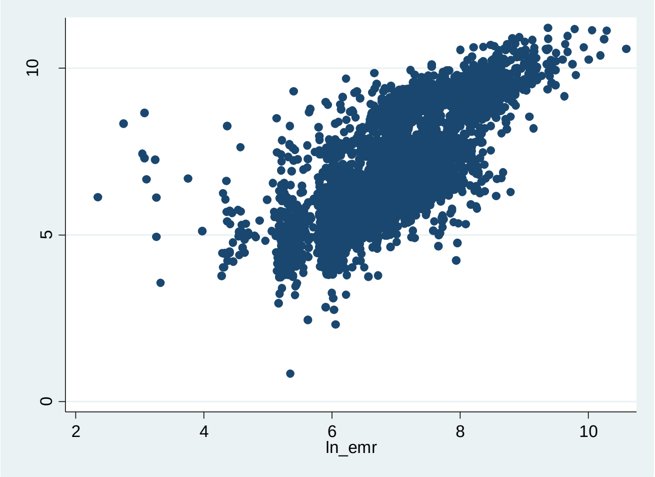
Results

(See the exhibits on the following page)



Correlation Coefficient: 0.81

Figure 1: Scatterplot of Estimated Cost of Care (All Inclusive)



Correlation Coefficient: 0.76

Figure 2: Scatterplot of Estimated Cost of Care (RA-related Only)

Figures 1 and 2 show the scatterplots of the log-transformed cost estimates as obtained from EMR and claims. Figure 2 is similar to Figure 1, except that only RA-related care (i.e., EMR and claims records for which ICD-9 code for RA appeared as one of the diagnosis codes) was considered. Each dot in the plots represents each patient in the sample. On the X-axis are the log-transformed 6-month cost estimates as obtained from EMR, and on the Y-axis are the log-transformed 6-month cost estimates as obtained from the claims data. The cost estimates were log-transformed to account for outliers and for enhanced visual representation of the data. The corresponding correlation coefficients in Figure 1 and Figure 2, respectively, are 0.81 and 0.76, suggesting relatively close agreement between EMR and claims data in terms of overall costs.

Table 1: Comparison of Cost and Utilization per 6-Month Period (All Inclusive)

	EMR (95% Bootstrap CI)	**Claims (95% Bootstrap CI)**	
Cost	$3,520 ($3,358 , $3,682)	$5,710 ($5,298 , $6,121)	*
ED	4.0% (3.4% , 4.5%)	11.2% (10.2% , 12.2%)	*
CT	5.1% (4.4% , 5.7%)	7.3% (6.4% , 8.1%)	*
MRI	4.8% (4.3% , 5.4%)	4.7% (4.0% , 5.3%)	
X-Ray	4.0% (3.6% , 4.5%)	22.0% (20.5% , 23.5%)	*
Inpatient	10.8% (9.9% , 11.8%)	11.4% (10.5% , 12.3%)	
Biologics	7.2% (6.1% , 8.3%)	6.7% (5.6% , 7.9%)	
DMARD	28.6% (26.6% , 30.7%)	30.5% (27.9% , 33.0%)	

* Statistically significant difference at 5% level

Table 2: Comparison of Cost and Utilization per 6-Month Period (RA-related Only)

	EMR (95% Bootstrap CI)	**Claims (95% Bootstrap CI)**	
Cost	$1,436 ($1,346 , $1,525)	$2,746 ($2,482 , $3,009)	*
ED	0.3% (0.1% , 0.4%)	2.3% (1.8% , 2.7%)	*
CT	1.0% (0.7% , 1.2%)	1.6% (1.2% , 2.0%)	
MRI	0.6% (0.4% , 0.9%)	0.8% (0.5% , 1.0%)	
X-Ray	2.1% (1.7% , 2.5%)	12.0% (10.8% , 13.2%)	*
Inpatient	6.9% (6.0% , 7.8%)	7.1% (6.2% , 7.9%)	
Biologics	5.4% (4.4% , 6.4%)	4.1% (3.2% , 5.1%)	
DMARD	30.0% (27.6% , 32.4%)	21.0% (18.9% , 23.1%)	*

* Statistically significant difference at 5% level

Tables 1 and 2 compare the mean total cost estimates as well as utilization patterns as obtained from the EMR and claims data. The estimates in Table 1 reflect all care and cost incurred (not just RA-related) during a given 6-month period per patient; Table 2 focuses only on RA-related care costs. Bootstrapped 95% confidence intervals are shown in all cases. Table 1 suggests that EMR-based estimated means of total cost of care and utilizations tend to underestimate the total cost of care. In particular, EMR understate ED visits (4% of the patients visited ED at least once during a six-month period vs. 11.2% in claims), X-rays (4% in EMR vs. 22% in claims), and CT scans (5.1% vs. 7.3%). Use of biologic agents appear to be slightly higher in EMR than in claims (7.2% vs. 6.7), although the difference is not statistically significant. A similar pattern is observed in Table 2.

Discussion

The results suggest that there is a lack of agreement between EMR and claims data in the study cohort. The scatterplots reveal that the degree to which these two data sets agree with each other is lower if only the care related to a specific condition (RA in this case) is considered. The main source of such discrepancies between EMR and claims appears to be missing utilization of certain types of care in EMR. In particular, discrepancies seem greater for the types of services for which patients have more alternative choices in the area (e.g. patients can visit EDs owned by non-GHS providers). Use of biologic agents appear to be slightly higher in EMR than in claims, most likely reflecting GHP's pre-authorization requirements; that is, some orders for biologic agents might have been denied by GHP.

There are several reasons why such discrepancies exist and are inherent in these data sources: First, the accuracy of the cost imputation method as presented above relies on subjective categorization of claim and encounter types. Second, claim data reflects health plan's coverage decisions and utilization management (e.g., pre-authorization requirements), while EMR reflects clinicians' decisions and practice patterns. Thus, researchers should carefully consider which "reality" they are interested in capturing in their analysis. Lastly, the fact that both EMR and claims are collected for clinical and administrative purposes, not for research purposes, must be emphasized. For instance, it should be recognized that the date information typically available in EMR and claims fundamentally capture two different timing of care: Dates in EMR typically represent the time at which the patient-provider encounter had taken place. Dates in the claims data, on the other hand, typically represent when the claim for the care was filed with the payer, not necessarily when the clinician had ordered care and when this care was actually provided.

On the other hand, more recent studies that have compared patterns of preventive care utilization using EMR and claims data[13,14] indicate a reasonable degree of agreement between the two data sources. This may be due to the fact that the identification of preventive services, which are major components of healthcare quality metrics such as Healthcare Effectiveness Data and Information Set (HEDIS), is more routine and standardized than it is for other types of care that are not necessarily preventive in nature. To the extent that accurate measurement of healthcare provider performance depends on routine and comprehensive data captures of all types of healthcare utilizations, further research is needed to improve the accuracy and reliability of using administratively collected EMR and claims as the standard data sources.

Study Funding: This study was funded by Janssen Pharmaceuticals, Inc., Titusville, NJ.

Acknowledgements: The study team gratefully acknowledges the input and consultations of Michael Ingham, from Janssen Scientific Affairs, Health Economics and Outcomes Research Division of Janssen Pharmaceuticals, Inc.

Reference

1. Tessier-Sherman B, Galusha D, Taiwo OA, Cantley L, Slade MD, Kirsche SR, Cullen MR. Further validation that claims data are a useful tool for epidemiologic research on hypertension. BMC Public Health. 2013 Jan 18;13:51.

2. Tang PC, Ralston M, Arrigotti MF, Qureshi L, Graham J. Comparison of Methodologies for Calculating Quality Measures Based on Administrative Data versus Clinical Data from an Electronic Health Record System: Implications for Performance Measures. J Am Med Inform Assoc. 2007 Jan-Feb;14(1):10-15.

3. Pawlson LG, Scholle SH, Powers A. Comparison of administrative-only versus administrative plus chart review data for reporting HEDIS hybrid measures. Am J Manag Care. 2007 Oct;13(10):553-558.

4. Liang SY, Phillips KA, Wang G, Keohane C, Armstrong J, Morris WM, Haas JS. Tradeoffs of using administrative claims and medical records to identify the use of personalized medicine for patients with breast cancer. Med Care. 2011 Jun;49(6):e1-8.

5. Hornbrook MC, Hart G, Ellis JL, Bachman DJ, Ansell G, Greene SM, Wagner EH, Pardee R, Schmidt MM, Geiger A, Butani AL, Field T, Fouayzi H, Miroshnik I, Liu L, Diseker R, Wells K, Krajenta R, Lamerato L, Neslund Dudas C. Building a virtual cancer research organization. J Natl Cancer Inst Monogr 2005(35):12-25.

6. Selby JV. Linking automated databases for research in managed care settings. Ann Intern Med 1997; 127 (8 Pt 2): 719-724.

7. Visser H, Cessie S, Vos K, Breedveld FC, Hazes JMW. How to diagnose rheumatoid arthritis early. Arthritis and Rheumatism. 2002: 46: 357-365.

8. Mottonen TT. Prediction of Erosiveness and Rate of Development of New Erosions in Early Rheumatoid Arthritis. Annals of the Rheumatic Diseases. 1988; 47: 648-653.

9. Curtis JR, Baddley JW, Yang S, Patkar N, Chen L, Delzell E, Mikuls TR, Saag KG, Singh J, Safford M, Cannon GW. Derivation and preliminary validation of an administrative claims-based algorithm for the effectiveness of medications for rheumatoid arthritis. Arthritis Res Ther. 2011; 13(5):R155.

10. Brian L. Strom, Stephen E Kimmel, Sean Hennessy. Pharmacoepidemiology, Fifth Edition. Hoboken, NJ: John Wiley-Blackwell, 2012.

11. Maeng DD, Stewart WF, Yan X, Boscarino JA, Mardekian J, Harnett J, Von Korff MR. Use of electronic health records for early detection of high-cost, low back pain patients. Pain Res Manag. 2015 Sep-Oct;20(5):234-40.

12. Stewart WF, Yan X, Boscarino JA, Maeng DD, Mardekian J, Sanchez RJ, Von Korff MR. Patterns of health care utilization for low back pain. J Pain Res. 2015 Aug 12;8:523-35

13. Heintzman J, Bailey SR, Hoopes MJ, Le T, Gold R, O'Malley JP, Cowburn S, Marino M, Krist A, DeVoe JE. Agreement of Medicaid claims and electronic health records for assessing preventive care quality among adults. J Am Med Inform Assoc. 2014 Jul-Aug;21(4):720-4.

14. Bailey SR, Heintzman JD, Marino M, Hoopes MJ, Hatch BA, Gold R, Cowburn SC, Nelson CA, Angier HE, DeVoe JE. Measuring Preventive Care Delivery: Comparing Rates Across Three Data Sources. Am J Prev Med. 2016 Nov;51(5):752-761.

Below is contact information for the authors:

Jing Hao, Ph.D
Assistant Professor
Department of Epidemiology and Health Services Research
Geisinger Health System
Danville, Pennsylvania USA
Email: jhao@geisinger.edu

Daniel D. Maeng, Ph.D (Corrsponding Author)
Assistant Professor
Department of Epidemiology and Health Services Research
Geisinger Health System
100 N. Academy Ave. M.C. 44-00
Danville, PA 17822 USA
Phone: 570-214-1688
Email: ddmaeng@geisinger.edu

Xiaowei Yan, PhD
Statistician Investigator
Sutter Health
Walnut Creek, California USA
Email: YanSX@sutterhealth.org

Walter F. Stewart, Ph.D, MPH
Chief Research Officer
Sutter Health
Walnut Creek, California USA
Email: StewarWF@sutterhealth.org

Joseph A. Boscarino, PhD, MPH
Professor
Department of Epidemiology and Health Services Research
Geisinger Health System
Danville, Pennsylvania USA
Email: jaboscarino@geisinger.edu