IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 5, 2019, accepted May 22, 2019, date of publication May 30, 2019, date of current version June 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919998

Estimation of Key Comorbidities for Osteoarthritis Progression Based on the EMR-Claims Dataset

WEI CHEN[1], KUN WEI[1], WEILING ZHAO[2], AND XIAOBO ZHOU[1,2]

[1]Center for Bioinformatics and Systems Biology, Division of Radiological Sciences, Wake Forest Baptist Medical Center, Winston-Salem, NC 27127, USA
[2]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Corresponding author: Xiaobo Zhou (xiaobo.zhou@uth.tmc.edu)

This work was supported by the Patient-Centered Outcomes Research Institute (PCORI), USA, through the project, Linking Electronic Medical Records to Medicare Claims Data, under Grant IRB00030804.

ABSTRACT Currently, some serious comorbidity-impacted chronic diseases have high incidence among older people in the U.S. Due to the incompleteness of the related clinical data, it is difficult to refine these diseases' staging and quantitatively assess risk effects of comorbidities on symptom progressions. Here, we used both electronic medical records (EMRs) and claims data to obtain a comprehensive data source in this paper. We adopted osteoarthritis (OA) as a demonstrated major disease. The key comorbidities and their risks for various OA stage-related progressions were estimated. We utilized the linked EMR-claims dataset of OA from 2007 to 2014. The EMR data provided pain scores and laboratory data, while claims data provided costs as a proxy for disease severity. Although both datasets contained diagnoses, procedures, and medications, the linked dataset included more distinct codes. We established a prototype to combine our developed relational dependency network (RDN) approach with Cox proportional models to extract and estimate key comorbidities' impacts on OA progression. We identified the key OA stage-related comorbidities. Our studies indicate that the combination of the EMR with claims data is a useful strategy for obtaining more accurate medical data sources from patients. The analyses of the impact of clinical factors on OA staging clarify the associations between key covariates and OA progression. These approaches can be generalized to summarize the impact of comorbidities on the development of various chronic diseases.

INDEX TERMS Osteoarthritis, comorbidity, linked EMR-claims dataset, disease stage, risk estimation, hazard ratio.

I. INTRODUCTION

Osteoarthritis (OA) is a degenerative joint disease, affecting over 30 million adults in the U.S [1]. In recent years, 33.6% of the population aged 65 and up in the United States have various types of OA [2]. Thus, there is an urgent need to develop effective mechanisms for the prevention and treatment of OA.

Some coexisting clinical conditions (i.e., diagnostic diseases, symptoms) have been known to play an important role in the development of OA [3]. The coexistent clinical factors condition is defined as comorbidities of OA [4]. Currently, most comorbidity-related OA studies focus on assessing the association between superficial OA symptoms and co-existing diseases [5]–[8]. Few studies have devoted to exploring the impact of comorbidities on OA progressions. Based on the etiopathogenesis, OA's progression has been divided into 3 sequential stages, including mild, moderate and severe stages [9]. OA symptoms for various stages may be caused by specific comorbidities [10]–[12]. It is necessary to quantitatively estimate comorbidities' impacts.

The clinical data that we used in this study include the electronic medical records (EMR) in Wake Forest Translational Data Warehouse (TDW) and the claims dataset purchased from Centers for Medicare & Medicaid Services (CMS) [13]. TDW EMR data is an important resource for studying clinical oncology, epidemiology, and comparative effectiveness. However, it is difficult to share TDW EMR data across different institutions due to the privacy rule. Claims dataset contains cross-institutional information, whereas the included clinical factors are not enough for clarifying the pathogenesis of diseases. Here, we combined TDW EMR data with the claims data to overcome limitations for using either claims or TDW EMR dataset alone.

The associate editor coordinating the review of this manuscript and approving it for publication was Yonghong Peng.

We adopted the relational dependency networks (RDN) approach [14] to extract the core clinical factors related to various OA stages. RDN reflects the dependency direction and strength among relational clinical instances [15]. It can be applied to association analyses of numerous biomarkers and clinical factors, which are represented as instance elements. Biomedical big data have provided tremendous data sources for pathologic and clinical knowledge discovery. RDNs obtained by various association approaches can delineate different knowledge networks, which reflect disparate subcategories of clinical studies. In order to handle poor data quality issues such as missing values and heterogeneity [16], we developed a novel RDN approach, named bootstrapping for unified feature association measurement (BUFAM) [17]. By applying BUFAM-RDN to the linked EMR-claims dataset, core clinical factors and association networks were derived. In addition, the relationship network among these extracted clinical factors was generated. We then did a risk analysis on these clinical factors and extracted the high-risk factors.

In our study, a prototype was established to extract and analyze key comorbidities' impacts on the stage progressions of the main disease. We combined our proposed association approach (BUFAM, bootstrapping for unified feature association measurement) with the classical risk estimation methodology (logistic, Cox proportional regression, and generalized linear model). The total pipeline is composed of 3 steps: (1) Construct Relational Dependency Network (RDN) through applying BUFAM to high-frequent clinical factors; (2) Based on the pathologic stage-related RDNs, generate new cox proportional regression and generalized linear models to summarize the high-frequent comorbidities' impacts (hazards and economic effects) on main disease's stage progression for collected patient records (training samples); (3) Predict the risks of stage changes and estimate disease stages for new patient records (testing samples). The 3 steps can be considered as a prototype to extract and analyze key comorbidities' impacts on the phase change of the main disease. We used Osteoarthritis (OA) as a demonstration case to estimate core comorbidities' effects. We have compared the results of our pipeline with those of traditional risk estimations on high-frequent clinical indices. It is evident that our pipeline can give new and reasonable discoveries related to comorbidities' effects. The prototype can be applied to any main disease with evident pathologic phases. For related physicians, our prediction models can provide the quantitative measurement of core symptoms' impacts on the stage progression of the main disease. Such estimation may guide clinicians' diagnosis and therapy decisions for personalized medicine.

II. MATERIALS AND METHODS

A. OVERVIEW OF EMR AND CLAIMS DATASET FOR OA

1) DATA SOURCE

The EMR data used in this study was from the i2b2-framework-based TDW [18], including 5 tables depicting demographics, clinical visit records, detailed symptom and therapy notes, medical concepts and provider information. Medical concepts refer to detailed clinical terminologies related to diagnoses, procedures, medications and lab tests.

The claims data was purchased from CMS and contained OA cases collected during 2007∼2014. The original claims dataset [19] consisted of 27 tables belonging to 5 categories, including 14 tables for the Healthcare providers, 4 for the summarization of beneficiaries, 1 for the prescription drugs, 2 for durable medical equipment and 6 for hospices. Multiple healthcare institutes were covered in the claims notes.

2) STUDY COHORT

We collected OA patient data from 1,300,000 patients' TDW EMR data (2001 - 2014) and 101,000 patients' claims data (2007 - 2014) from the Wake Forest Baptist Medical Center and Translational Science Institute (CTSI).

TDW EMR data is only from one provider (Wake Forest Baptist Medical Center), while claims data is from multiple healthcare institutions. For lab test records, we can have access to the longitudinal notes in one institution during the censoring period (records over time). If the patient moves to another provider, we can trace his/her information of diagnoses, therapies, and medications. In general, the linked EMR-claims data can provide integral clinical information related to the specific patient over the censoring period.

In order to construct the linked EMR-claims dataset, we screened OA patients based on 2 prerequisites, including that (1) clinical records must be in both of the TDW EMR and claims datasets, and (2) the OA diagnosis records should appear at least once either in TDW EMR or the claims dataset during 2007 ∼ 2014. We determined the cohorts by analyzing TDW EMR and claims data independently and then extracted 8,480 patients with OA records in TDW EMR and claims data. During our censoring period (2007 ∼ 2014), 20,614 patients in TDW EMR had diagnosis records of OA (including different subcategories of OA such as knee OA, hip OA), while 8,480 patients had claims data. In addition, we obtained the patient ID mapping table between TDW EMR and claims data from Wake Forest Clinical and Translational Science Institute (CTSI).

Finally, the EMR and claims records of these 8,480 patients during 2007 ∼ 2014 were selected as our study cohort.

3) COVARIATES

Covariates refer to various medical terminologies, which own coded values reflecting the terminology-related medical concept and appear in Table "clinical visit records" and "symptom & therapy notes". Table "Medical concepts" stores a mapping table between specific clinical terminologies and related coded values.

We extracted all TDW EMR and claims records for each patient. The information of TDW EMR included demographic, clinical visits (diagnosis), medical observation (drugs, clinical procedures) and detailed providers.

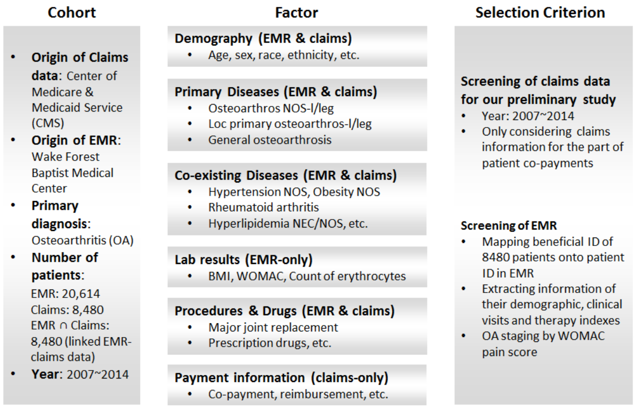

Cohort
Factor
Selection Criterion
Origin of Claims data: Center of Medicare & Medicaid Service (CMS)
Origin of EMR: Wake Forest Baptist Medical Center
Primary diagnosis: Osteoarthritis (OA)
Number of patients: EMR: 20,614 Claims: 8,480 EMR ∩ Claims: 8,480 (linked EMR-claims data)
Year: 2007~2014
Demography (EMR & claims)
Age, sex, race, ethnicity, etc.
Primary Diseases (EMR & claims)
Osteoarthros NOS-l/leg
Loc primary osteoarthros-l/leg
General osteoarthrosis
Co-existing Diseases (EMR & claims)
Hypertension NOS, Obesity NOS
Rheumatoid arthritis
Hyperlipidemia NEC/NOS, etc.
Lab results (EMR-only)
BMI, WOMAC, Count of erythrocytes
Procedures & Drugs (EMR & claims)
Major joint replacement
Prescription drugs, etc.
Payment information (claims-only)
Co-payment, reimbursement, etc.
Screening of claims data for our preliminary study
Year: 2007~2014
Only considering claims information for the part of patient co-payments
Screening of EMR
Mapping beneficial ID of 8480 patients onto patient ID in EMR
Extracting information of their demographic, clinical visits and therapy indexes
OA staging by WOMAC pain score

FIGURE 1. An overview of the linked EMR-claims dataset for Osteoarthritis. Data cohorts refer to medicare & medicaid claims dataset from multiple healthcare institutes and TDW EMR from wake forest baptist medical center. A total of 8,480 patients with OA records were included in the study. The censoring period was from 2007 to 2014. Major clinical factors and screening criteria are listed.

The contents of claims data involved the medical costs, demographics, diagnosis, dispensing, procedures, encounter, and enrollment. OA symptoms were classified into leg OA, localized primary OA, and general OA, as shown in Figure1. The year 2007 was used as the baseline.

In Figure 1, for the factors from both sources, we adopted the union set of clinical records. TDW EMR from Wake Forest Baptist Medical Center can provide patients' detailed clinical diagnostic symptoms or disease information, while claims data can give all of the patients' payment-related diagnosis information and diagnostic records in various medical providers. Patients' disease status can be fully reflected by combining TDW EMR and claims data.

For each of 8,480 OA patients, we have searched his/her longitudinal records of the merged EMR-claims dataset from 2007 to 2014. For diagnoses, TDW EMR's "Clinical_visit" table and claim's "Diagnosis" table provide detailed information with ICD-9 codes. On average, there are 43 unique ICD-9 codes (variables) for each patient with a standard deviation of 31. So there exist large variances of diagnosis among different patients. The procedure Information is included in TDW EMR's table with ICD-9, HCPCS and CPT codes. There is an average of 19 unique procedures (variables) for each patient. The standard deviation is 13. Drug records are from TDW EMR's "observation_fact" table and claim's "dispensing" table. According to the mapped annotations, there exist 16 unique drugs (variables) with a standard deviation of 28. Thus, some patients accept more drug therapies than others.

In each year of 2007 ∼ 2014, all of 8,480 OA patients have clinical visits at Wake Forest Baptist Medical Center (shown in both TDW EMR and claims data) or other healthcare providers (shown in claims data). Based on TDW EMR's "Clinical_visit" table and claim's "Encounter" table, each patient had 14 (mean) ± 5 (std) clinical visits per year. The general statisticsare shown in Table 1.

TABLE 1. Clinical factors number of diagnosis, procedure, drug and lab results in TDW EMR, claims and linked EMR-claims data for 8,480 OA patients during the censoring period (2007 ∼ 2014).

Type	Number in distinct TDW EMR data	Number in distinct claims data	Overlapped number	Number in linked EMR-claims data
Diagnosis	3,882	1,124	1,151	6,157
Procedure	1,344	815	154	2,313
Drug	543	297	84	924
Lab result	163	Nil	0	163

4) OUTCOMES

The primary outcome was OA-related diagnosis records (EMR and claims data), WOMAC pain scores (EMR) and various medical expenditures (claims data). The three outcomes are applied to the extraction of OA patients, OA staging and estimations of comorbidities' dynamic impacts on OA progressions.

5) STATISTICAL ANALYSIS

In order to extract reasonable clinical indices related to OA stages, we applied RDN to original medical records. RDN covered various association test approaches such as Spearman-test, One-way test, and Pearson correlation. Regarding the estimation of OA stages, we executed the logistic regression on candidate associated clinical characteristics. In addition, the Cox proportional regression was applied to the prediction of OA progression risks. Finally, we used the generalized linear model to both expenditure and clinical records for estimating the dynamic impacts of OA's comorbidities.

B. STAGING OF OA

The patients were first classified into mild, moderate and severe stages of OA using the Western Ontario and McMaster Universities Arthritis Index (WOMAC) pain score, a standard measurement tool for knee and hip OA [20].

WOMAC scores are from the self-administered questionnaire consisting of 24 items for 3 subscales, including pain, stiffness and physical functions. The information of questionnaires is stored in EPIC system of Wake Forest Medical School. All contents in TDW EMR data are from Wake EPIC system. We extracted the questionnaire records from EPIC and summed the scores related to 3 subscale factors. The summed scores were considered as the WOMAC scores. With reference to the initial OA staging of individual patients, we calculated the average WOMAC score for each patient during the whole censoring period (2007 ∼ 2014). Based on the WOMAC scoring threshold, each patient was classified into a specific OA stage. The WOMAC scores between 0 and 1 represented two conditions at the mild stage (None and slight); the scores within 1 ∼ 3 for the moderate stage and those within 3 ∼ 4 for the severe stage (very and extremely). Of 8,480 OA patients, 4,313 patients were in the mild stage, 2,476 in the moderate stage, and 1,691 in the severe stage.

Here, the average of annual WOMAC scores is a general approach for staging OA patients. Based on the limited WOMAC records in our TDW EMR data, the average of patients' WOMAC scores during the whole censoring period can generate the comparatively precise groups of patients related to 3 OA stages. If we only utilize WOMAC notes in shorter periods or adopt separate WOMAC information on specific dates, patients' OA staging will be less accurate. In addition, the patient number will be reduced because of the missing issue of WOMAC information in shorter periods.

C. CONSTRUCTION OF THE RDN GRAPH RELATED TO SPECIFIC OA STAGE

In OA-related study, we developed BUFAM (Bootstrapping for unified feature association measurement), which is an association analysis approach for pairwise clinical factors. BUFAM provides a uniform metric to enable an unbiased analysis of pairwise feature associations across datasets. Based on the different combinations of data types (numeric, binary, categorical, ordinal) related to clinical factors, BUFAM will select different statistical measurements (e.g., spearman, one-way, Chi-square test) to derive the association weights accurately. Furthermore, BUFAM can (1) deal with the data heterogeneity issue; (2) solve the multi-site harmonization; and (3) address the missing data problem [21]. Thus, BUFAM enables efficient and effective use of heterogeneous biomedical big data for reflecting clinical factors' associations and understanding diseases.

Based on the frequency distribution of clinical covariates, we selected 200 of the most frequently occurring factors and calculated the average value of each patient during the censoring period (2007-2014). In general, numeric and ordinal factors are represented by algebra; while binary and sub-indexes of categorical factors are shown as occurrences. For a specific clinical covariate, if one patient does not share the related information, the element value of covariate-patient is NA. Thus, the matrix with the most frequent clinical factors was constructed. We then applied BUFAM-RDN to the matrix to obtain the pairwise associations and calculated the *p*-values between any two specific clinical factors. Based on the *p*-value threshold (0.01) and the presumed values of in-degree and out-degree for nodes (7), the significantly associated pairs were used as the nodes and edges in the final RDN graph.

Within the linked EMR-claims dataset, rare (low-frequency) disease indices coexisting with OA are shared by few patients. There exist large numbers of NA values related to most patients, which make it impossible to apply BUFAM for constructing RDN. For choosing 200 most frequent features, we have checked the status of missing values in our dataset and observed that the missing data cases within 200 most frequent clinical factors can be handled by BUFAM. If more features are included, BUFAM will give less reliable estimations. We have tried other approaches such as the causality analysis (Partial causality, Granger causality). However, the missing data problem and multi-site harmonization cannot be solved well. BUFAM is more adaptive to the dataset with heterogeneity and missing record issues.

D. ESTIMATION OF OA STAGES USING THE LOGISTIC REGRESSION MODEL

WOMAC pain score has been used for the OA stage identification [22]. However, it is often inconvenient to obtain the related integral information [23]. This reflects the need to estimate the OA stages using non-WOMAC medical information. Since the WOMAC records are often incomplete, the OA stages cannot be accurately evaluated. However, we can apply a logistic regression model to the available clinical records to build a model for evaluating OA stages. Since the selected non-WOMAC clinical records are complete, we are able to construct the specific logistic regression model.

We estimated OA stages using clinical characteristics identified by RDNs. Since the data type of stage measurement is binary, we adopted the logistic regression model to estimate the probability of a particular OA stage. Three logistic models were constructed to estimate the likelihood of mild, moderate and severe stages, respectively. The input covariates are the demographic features and characteristics in the major modules of RDN. The output is a binary variable for identifying OA stage, which is derived from the index of pain measurement.

Generally, we utilized the standard logistic regression based on the principle of maximum likelihood. For model parameters, we guaranteed that the number of parameters was smaller than that of included patients (No. of coefficients $<$ No. of observations). The order of selecting coefficients (clinical factors) conformed to the decreasing order of their appearance frequencies. In addition, we used the standard Newton-Raphson algorithm to extract specific coefficients (clinical factors) maximizing the likelihood of the data. Finally, the trained logistic model contained these selected coefficients. For the calibration of our models, the standard logistic regression returns well-calibrated predictions by default as it directly optimizes log-loss, which means the automatic calibration to fit the training data. We divided the patients of each stage into derivation and validation cohorts using 10-fold cross-validation strategy (see Table 2). Then, the 10-fold results were averaged to produce final estimations. We estimated two types of evaluation indexes, including (1) observed ratio and predicted risk probability (Figure 3); and (2) sensitivity & specificity (Table 4)) The vertical axis of Figure 3 represents the portion of patients classified into a specific phase among all 3 stages. Figure 3 is composed of 3 parts related to 3 subcategories of OA. We used 3 logistic models to analyze the same patient's data and obtained 3 risk probabilities corresponding to the OA stages. The resulting highest risk probability was used to classify patients into a specific OA stage.

TABLE 2. The sample numbers used in the derivation and validation cohorts.

Stage	Derivation cohort	Validation cohort
Mild	3,881	432
Moderate	2,228	248
Severe	1,522	169

E. SUMMARIZATION OF COMORBIDITY HAZARDS ON OSTEOARTHRITIS PROGRESSION

Based on RDN-derived clinical factors, we utilized the logistic regression to construct the estimation models for different OA stages. For new patients, we can classify OA stages by using the trained models upon their clinical records. In addition, the OA patients in each stage group presented with different health states, such as deterioration, relief, and maintenance. Analyzing the impact tendencies of covariates on OA deterioration or relief can provide a basis for preventive and therapeutic strategies.

We defined 4 cases of OA stage changes, denoted mild → moderate as Case 1; moderate → severe as Case 2; moderate → mild as Case 3; and severe → moderate as Case 4. Based on the analysis of each patient's costs and OA-related clinical factors, the patients were attributed into 4 subgroups. The number of patients in 4 categories was 1527 (Case 1), 486 (Case 2), 1125 (Case 3) and 267 (Case 4), respectively. We also recorded the starting dates of the OA stage transformation of each patient for the purpose of risk estimation.

Here, for each of 8480 patients, we calculated the average WOMAC score for every year from 2007 to 2014. For the year without available WOMAC records, the average WOMAC score was displayed as NA. In addition, based on the mapping relationship between WOMAC score and OA stage, annual WOMAC scores were transformed into OA stages.

For the relationship between the OA stage and annual medical expenditures, we have found out two sources. In [8], the cost of treatment patterns, clinical comorbidities, and direct medical costs are summarized based on a retrospective claims database analysis. On the webpage of arthritis in Illinois bone & joint institution ([24], https://www.ibji.com/arthritis-in-knee-4-stages-of-osteoarthritis/), the detailed treatment strategies related to mild, moderate and severe stages are listed. Through combining the two sources, we consider that the cost of claims can roughly represent different stages of OA.

We then calculated the annual expenditure from 2007 to 2014. The cost range related to each OA stage was expressed as "Mean ± standard deviation". We acquired the expenditures as \$408.12 $\sim$ \$946.64, \$1844.49 $\sim$ \$2852.73, and \$4015.62 $\sim$ \$7280.78 for mild, moderate and severe OA patients, respectively. Based on the difference in the expenditure interval, we transformed annual costs into typical OA stages. Therefore, we obtained two tables with the dimension as 8480 patients $*$ 8 years, including (1) OA stage table by annual WOMAC (existing missing values) and (2) OA stage table by annual payment. We then extracted the elements with the same staging values in both tables. Based on the extracted regions corresponding to each patient, we identified the first consecutive year area and disease changing patterns such as mild → moderate, severe → severe. Then we attributed the related patients into 4 cases of stage changes (Mild → Moderate, Moderate → Severe, Moderate → Mild, Severe → Moderate). Here, two issues were involved in general. (1) Estimate OA stages for 8480 patients in each year during the censoring period (2007 $\sim$ 2014) based on annual average WOMAC scores and cost summations (Results: two OA stage tables (Row: patients, Column: years (2007 $\sim$ 2014))) (2) Based on the intersection regions in two OA stage tables, attribute related patients to 4 cases of OA stage progressions. For (1), we adopted the annual average WOMAC score and expenditure summation as two OA-stage mapping indices. Then, through the WOMAC via OA stage transformation (Section "Staging of OA") and annual cost intervals for different OA stages, we derived the specific OA stage related to each patient in each year. For (2), we picked up the common regions (patient via year) between two OA stage tables (one derived from WOMAC transformations, another from cost transformations). Then based on the identification of the first consecutive year area and the changing pattern of disease stages, we assigned each related patient to one of 4 stage change cases.

The Cox proportional regression model was employed to estimate the risk of OA stage changes. The outcomes (events) were four cases. Referring to the input covariates "duration time" and "risk factor", the duration time (until OA stage changes) was obtained through analyzing time series data of each patient, while the risk factors were the 10 most frequently co-occurring characteristics as shown in Table 3. Our monitored comorbidities for the source stages of 4 cases are originated from the high-frequent co-existing diseases or treatments in Table 3. For instance, in Case 1 (Mild OA $- >$ Moderate OA), the source stage is Mild. Then, the results related to Case 1 involve all 10 high-frequent comorbidities related to Mild stage in Table 3. We monitored the ten comorbidities' risks on the OA stage progression towards Moderate stage.

To reflect the dynamic effect of key comorbidities on OA progression, we applied a generalized linear model (GLM) to calculating covariates' coefficients for every year [25]. Here, "dynamic effect" means the tendency of comorbidities' impact changes among different years. It measures the trend of comorbidities' impacts on OA stage progressions. The outcome of GLM is the clinical expenditure, an indirect reflection of OA severity levels.

We calculated annual expenditures for each patient. The payments were involved in the diagnosis, procedure and dispensing drugs. For each patient, we summed all costs related to both self-burdened and insurance reimbursed portions in each year during the censoring period. The average

TABLE 3. 10 High-frequency occurrence covariates for various OA stages.

OA Stage	10 most frequently co-existing diseases or treatments with OA	
Mild	C1: Major joint replacement	C6: Periostitis unspecific
	C2: Hypertension NOS	C7: Furosemide
	C3: joint pain-l/leg	C8: Obesity NOS
	C4: Hypothyroidism NOS	C9: Edema
	C5: Rheumatoid arthritis	C10: Benign hypertension
Moderate	M1: Joint replaced knee	M6: Backache NOS
	M2: Joint pain pelvis	M7: Abnormal pain NOS
	M3: Lumbago	M8: Major joint replacement
	M4: Pure hypercholesterolem	M9: Malaise and fatigue
	M5: Hyperlipidemia NOS	M10: Warfarin sodium
Severe	S1: Joint replaced knee	S6: Furosemide
	S2: Hydrocodone acetamino	S7: Chronic kidney disease
	S3: Aftercare joint replace	S8: Chest pain NOS
	S4: Major joint replacement	S9: Backache NOS
	S5: Respiratory abnormal NEC	S10: Malaise and fatigue

annual total costs for the specific patient were then calculated. Based on the classified patient cohorts associated with 3 OA stages, we performed a Kruskal-Wallis test on the average annual expenditures for 3 groups of patients. "Kruskal-Wallis" test is a non-parametric approach for testing whether samples originate from the same distribution [26]. It can be used for comparing multiple independent sample sets with different sample sizes. Here, we adopt the "Kruskal-Wallis" test to check whether there exist evident differences among the expenditures related to patients for 3 OA stages. The returned p-value is 0.0426, which reflects that there exist evident differences among the expenditures for 3 OA stage-related patient groups. So there exist evident annual cost differences among the patients with different OA stages. Finally, we checked the annual payment range for each patient cohort; \$677.38 ± 269.26 for Mild-stage cohort, \$2348.61 ± 504.12 for Moderate-stage cohort, and \$5648.20 ± 1632.58 for Severe-stage cohort. Therefore, Severe OA patients spent more than mild and moderate OA patients. In our GLM, "comorbidity" is regarded as covariates associated with the change of OA stages. Figure 6 reflects the comorbidity impacts on OA stage changes. Here, we clarify how to apply comorbidity indices to our GLM. High-risk comorbidities derived by the Cox proportional model (shown in Figure 4) are regarded as input covariates in GLM. The coefficients related to comorbidities represent the association between the OA stage and specific comorbidities. This provides some quantitative clues for physicians to analyze the comorbidity-related OA stage transformation. We provided some co-existing diseases and symptoms as candidate factors. In this way, physicians can narrow down the clinical covariates affecting OA status progression. In addition, earlier studies [27], [28] have shown evidence that some diseases (e.g. Obesity, hypertension) aggravate the deterioration of OA stages. Our study provides a quantitative measurement of aggravation levels.

III. RESULTS

A. ESTABLISHMENT OF THE SPECIFIC NETWORKS FOR VARIOUS STAGES OF OA USING RDN APPROACHES

Three association network graphs for each OA stage were obtained. Four types of features are displayed in Figure 2, including diagnosis (circle), procedure (diamond), diagnosis-related group (DRG; hexagon) and dispensing (square). Figure 2A, 2B & 2C show RDN modules for mild, moderate and severe OA stage, respectively.

Through the network analysis of 3 OA stages, we narrowed down the key covariates. Here, our network analysis referred to the calculation of the in-degree and out-degree for each node. In Figure 2, the nodes with summation values of in-degree and out-degree larger than the specific thresholds were considered as the harbor nodes (Thresholds: Figure 2A & 2B: 7; Figure 2C: 5). In addition, we did the modularity analysis for the networks in Figure 2A and 2B. There existed 3 modules in both networks. We checked the harbor nodes belonging to the major module (including most nodes in the entire network) and summarized their clinical meanings and categories, which helped us to narrow down the key covariates. Based on these analyses, the blood-related diseases (i.e., Hypertension, hyperlipidemia) were common in the mild and moderate OA stages, while respiratory diseases & various pain symptoms were associated with the severe stage. In addition, the major modules of RDNs provided candidate covariates for risk estimations of OA stages and onset processes.

B. RISK ESTIMATION OF OA STAGE

We assessed the sensitivity and specificity of the estimated OA stages using logistic models. Since the output is binary, a measurement of performing a binary classification test is needed. Sensitivity and specificity analysis is such the statistical measurement. Thus, we adopted the sensitivity and specificity to evaluate the performance of logistic models. For a specific patient, we applied his/her clinical records to 3 trained logistic models for 3 OA stages. Then, 3 probability values were outputs related to 3 OA stages. The predicted probability threshold was 0.6 (empirical value). The OA stage with the probability value larger than 0.6 and the maximum value among 3 probabilities was assigned to this patient. Here, the threshold is an optional parameter for determining OA stages. The probabilities greater than the threshold can strengthen the reliability of OA stage predictions. Furthermore, the index "AUC" was also calculated. We firstly constructed the ROC curves for our 9 logistic classifiers related to 9 columns in Table 4. We set 10 discrimination thresholds for all samples and calculated the true positive rate (TPR) and false positive rate (FPR) related to each threshold. ROC curves were composed of these points (TPR & FPR). Then, AUCs were calculated related to 9 ROC curves. The related results were displayed in Table 4 and Figure 3. As shown in Table 4, the sensitivity and specificity are greater than 0.7 in 13 out of 18 testing items, indicating that the prediction from our model is reliable. However, we observe a large

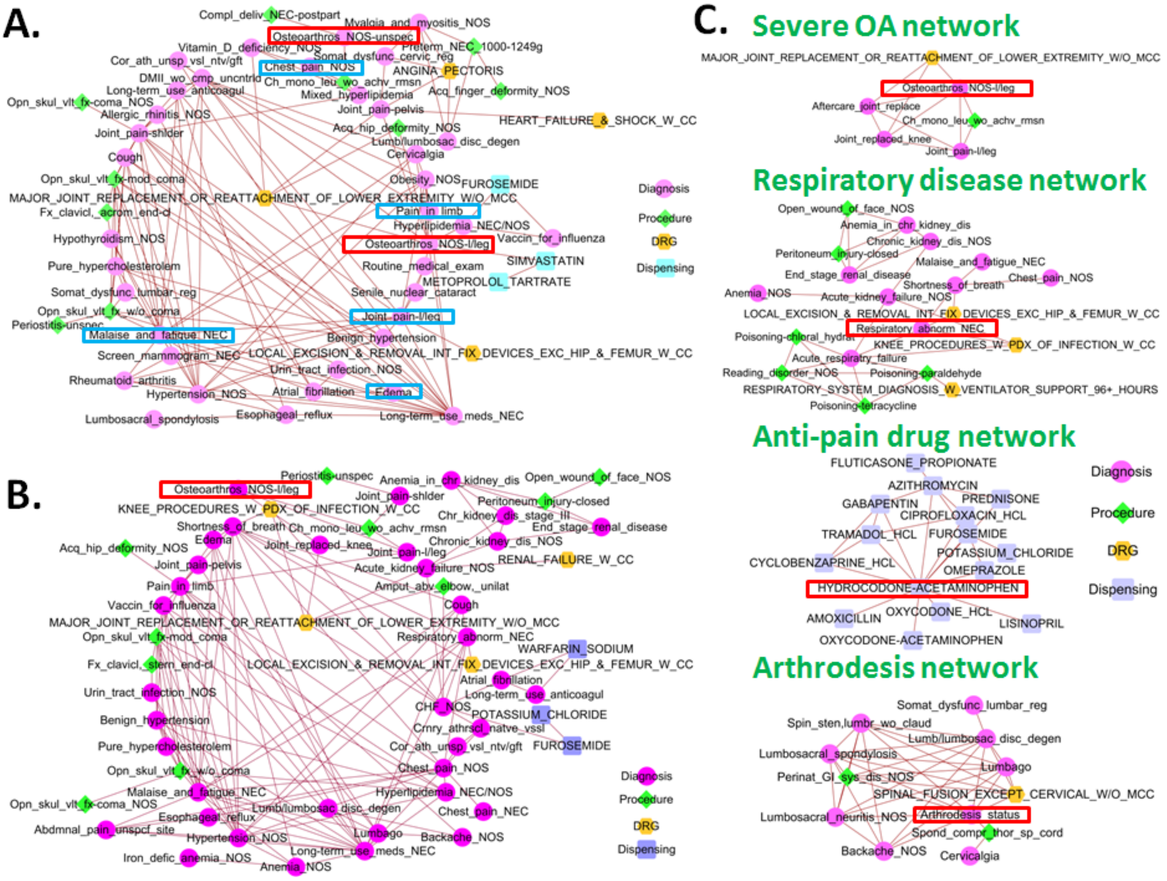


FIGURE 2. Major RDN modules for different stages of OA. (A). RDNs for the mild stage of OA, including 79 nodes and 147 links; (B). RDNs for the moderate stage of OA, covering 79 nodes and 147 links; (C). RDNs for the severe stage of OA, containing 4 networks: Severe OA, respiratory disease, anti-pain drug and arthrodesis. All 3 sub-figures involve 4 types of nodes (clinical factors) belonging to diagnosis, procedure, diagnosis-related group (DRG) and dispensing.

gap between the sensitivity and specificity values related to 3 OA cases (General OA Moderate, leg OA Mild, leg OA Moderate), indicating poor estimation performance. The likely cause is that the three OA cases have fewer training samples than other OA cases. In addition, RDN-derived candidate clinical factors may be less accurate due to the small sample status. With 9 OA cases showing in Table 4, logistic regressions can generally provide the reasonable estimation of OA stages. Figure 3 shows the comparison between the ratio of observed stage-related patient numbers and that of predicted ones. Seven out of nine groups have a deviation of less than 10%. A 15% deviation was observed for the groups with mild and moderate stages of leg OA. Our results indicate that the logistic regression model can reliably estimate OA stages.

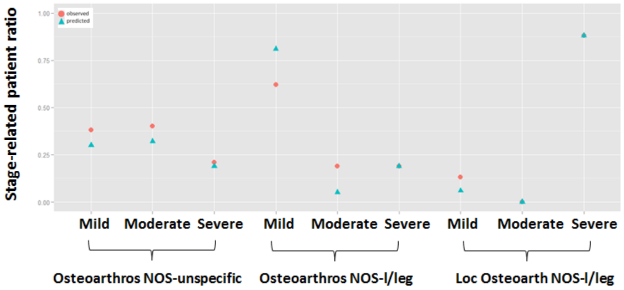


FIGURE 3. Ratios of observed and predicted patient numbers related to mild, moderate and severe stages for 3 subcategories of OA. Osteoarthrosis NOS-unspecific, Osteoarthrosis NOS-l/leg and localized Osteoarthrosis NOS-l/leg are 3 subcategories of Osteoarthritis (OA). The ratio of estimated patient numbers for mild, moderate and severe stages related to each subcategory of OA is calculated. Red points show the observed patient ratio, while green ones give the predicted ratios.

C. THE KEY COMORBIDITIES ASSOCIATED WITH OA PROGRESSION

Figure 4 shows the hazard ratios of various covariates for the derivation cohort for 4 cases. A hazard ratio larger than 1 represents a higher risk (positive association), while the hazard ratio less than 1 means a lower risk (inverse associations) [29]. The effects of various covariates on OA progression are listed in Figure 4. Red bars represent the high-risk covariates.

To validate the Cox proportional models, we grasped the patients in the validation cohort as described in

TABLE 4. Sensitivity & specificity analysis of the estimated OA stages.

	General Osteoarthritis			Osteoarthritis NOS-l/leg			Loc primary Osteoarthritis-l/leg		
OA stage	Mild	Moderate	Severe	Mild	Moderate	Severe	Mild	Moderate	Severe
Sensitivity	0.79	0.34	0.87	0.96	0.16	0.72	0.57	0.82	0.94
Specificity	0.83	0.71	0.85	0.37	0.94	0.88	0.84	0.92	0.50
AUC	0.78	0.64	0.82	0.66	0.57	0.76	0.72	0.84	0.73

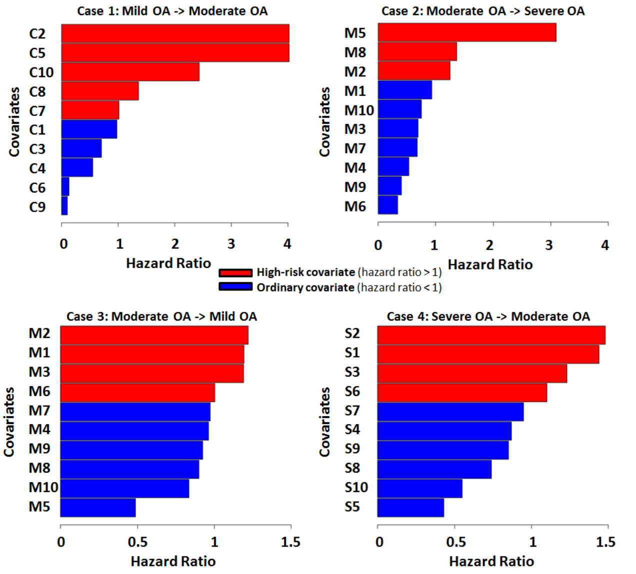


FIGURE 4. Covariate impacts (hazard ratios) on the deterioration & relief of OA stages. 4 cases of OA stage transformation are listed. For each case, the hazard ratios for 10 high-frequency occurrence covariates (in Table 3) are shown in the bar-plots. Red bars represent high-risk clinical factors, while blue bars show ordinary covariates.

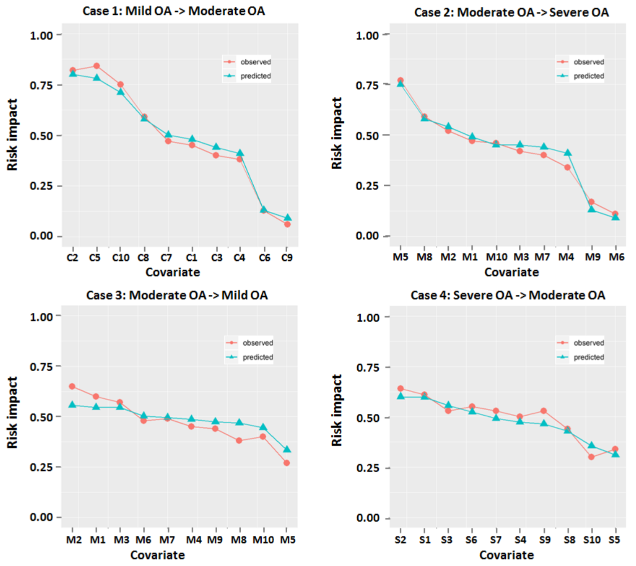


FIGURE 5. Observed and predicted probability of OA stage transformations. The observed risks of OA stage changes are acquired by K-M analysis, while the predicted probabilities of stage transformations are derived from hazard ratios. Comorbidity risks for 4 cases of OA stage transformations are shown. For each case, 10 high-frequency covariates are sorted based on predicted risk values. Green curves reflect predicted risks, while red ones give observed probabilities.

the Section "Summarization of comorbidity hazards on osteoarthritis progression". For each of 4 cases, we adopted the same covariates as the derivation cohort and applied the trained Cox proportional model to estimating the risks (risk = hazard/(1+hazard), range: 0 ∼ 1) (Each covariate [30] in the observed risks were calculated using Kaplan-Meier estimator [31].

The observed ratio is extracted from K-M analysis, while the predicted risk probability is calculated based on hazard ratios. The estimated and observed risks are shown in Figure 5. It is evident that the curves of predicted risk probabilities are close to those of observed ratios in all 4 cases. In addition, we calculate the average relative errors (Equation (1)) for 4 cases.

$$\Gamma = \frac{|\mathrm{Pr}_{risk} - \mathrm{Pr}_{observ}|}{\mathrm{Pr}_{observ}} \tag{1}$$

$\mathrm{Pr}_{\mathrm{risk}}$: predicted risk probability, $\mathrm{Pr}_{\mathrm{observ}}$: observed ratio.

Case 1: 3%, Case 2: 4%, Case 3: 5%, Case 4: 5%. Based on the curve comparison and relative errors, our prediction model is comparatively reliable.

D. DTHE DYNAMIC EFFECT OF KEY COMORBIDITIES ON OA STAGE PROGRESSION

Figure 6 shows the dynamic impact of high-risk covariates on OA progression during the censoring period (2007 ∼ 2014). Most factors have a progressive effect on OA, for example, furosemide (C7) and hypertension (C2) for the Case 1. Some of the factors show an oscillating trend, such as benign hypertension (C10) for the Case 1. The influence tendency provides important guidance for medical monitoring.

The observation values in each time trend refer to the weights of clinical factors (comorbidities) in the generalized linear model (GLM). These clinical characteristics with high weight are high-risk comorbidities (HR > 1, red bars) as shown in Figure 4. Since all four cases in Figure 4 have OA stage transformation status nearby the year 2014 (end of the censoring period), the high-risk comorbidities may generate a significant impact around 2014. The results in Figure 6 displays such dynamic impact trends.

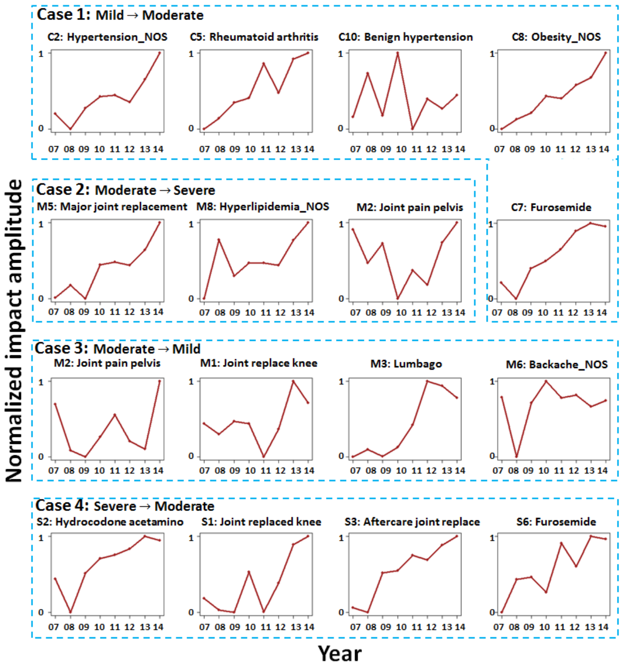
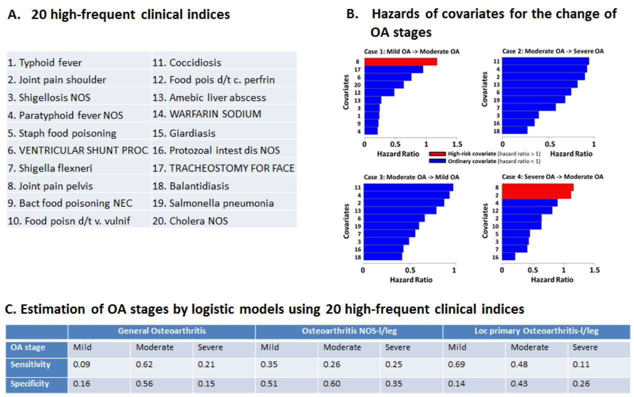


FIGURE 6. Tendency tracking of key comorbidities' impacts during the censoring period (2007 ∼ 2014).


A. 20 high-frequent clinical indices

1. Typhoid fever	11. Coccidiosis
2. Joint pain shoulder	12. Food pois d/t c. perfrin
3. Shigellosis NOS	13. Amebic liver abscess
4. Paratyphoid fever NOS	14. WARFARIN SODIUM
5. Staph food poisoning	15. Giardiasis
6. VENTRICULAR SHUNT PROC	16. Protozoal intest dis NOS
7. Shigella flexneri	17. TRACHEOSTOMY FOR FACE
8. Joint pain pelvis	18. Balantidiasis
9. Bact food poisoning NEC	19. Salmonella pneumonia
10. Food poisn d/t v. vulnif	20. Cholera NOS

B. Hazards of covariates for the change of OA stages

Case 1: Mild OA -> Moderate OA
8 17 6 20 12 13 3 1 9 4
Case 2: Moderate OA -> Severe OA
11 4 2 13 6 19 7 3 16 18
Case 3: Moderate OA -> Mild OA
11 4 2 13 6 19 7 3 16 18
Case 4: Severe OA -> Moderate OA
8 2 4 12 2 10 5 3 7 16
Covariates
Hazard Ratio
0 0.5 1 1.5
High-risk covariate (hazard ratio > 1)
Ordinary covariate (hazard ratio < 1)

C. Estimation of OA stages by logistic models using 20 high-frequent clinical indices

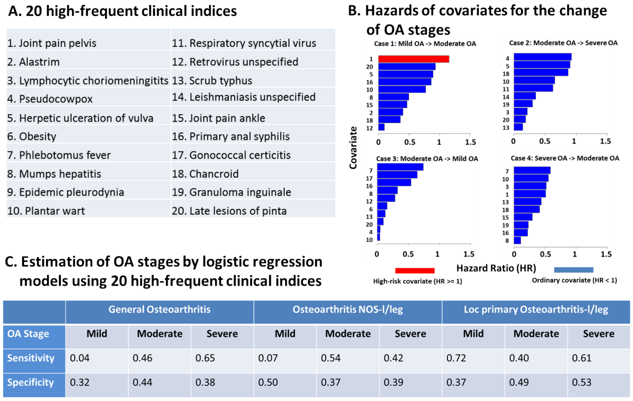
	General Osteoarthritis			Osteoarthritis NOS-l/leg			Loc primary Osteoarthritis-l/leg		
OA stage	Mild	Moderate	Severe	Mild	Moderate	Severe	Mild	Moderate	Severe
Sensitivity	0.09	0.62	0.21	0.35	0.26	0.25	0.69	0.48	0.11
Specificity	0.16	0.56	0.15	0.51	0.60	0.35	0.14	0.43	0.26

FIGURE 7. Risk estimation of OA stages and comorbidities without feature selection. (A). 20 high-frequent clinical indices generated without RDN; (B). Covariate hazards for OA stage progressions; (C). Estimation of OA stages by logistic regression models upon 20 high-frequent clinical indices.

E. RISK ESTIMATION OF COMORBIDITIES WITHOUT RDN

Regarding the effect evaluation of RDN, we applied partial correlation [32] and adjusted cosine similarity [33], [34] to the feature selection. Meanwhile, the direct utilization of original clinical features was also adopted. After that, Cox proportional and logistic regressions were applied to the comorbidity estimation.

Figure 7, 8 and 9 show the comorbidity risk estimation results obtained without feature selection, partial correlation, and adjusted cosine similarity, respectively. Compared to Table 3 (RDN-derived clinical factors), the panel A in Figure 7, 8 and 9 gives more redundant candidate clinical indices associated with OA. In addition, the clinical concept categories only refer to "diagnosis". There are no "procedure", "DRG" and "dispensing medication" indicating that these feature selection methods lead to the biased acquisition of clinical information.


A. 20 high-frequent clinical indices

1. Joint pain pelvis	11. Respiratory syncytial virus
2. Alastrim	12. Retrovirus unspecified
3. Lymphocytic choriomeningitits	13. Scrub typhus
4. Pseudocowpox	14. Leishmaniasis unspecified
5. Herpetic ulceration of vulva	15. Joint pain ankle
6. Obesity	16. Primary anal syphilis
7. Phlebotomus fever	17. Gonococcal certicitis
8. Mumps hepatitis	18. Chancroid
9. Epidemic pleurodynia	19. Granuloma inguinale
10. Plantar wart	20. Late lesions of pinta

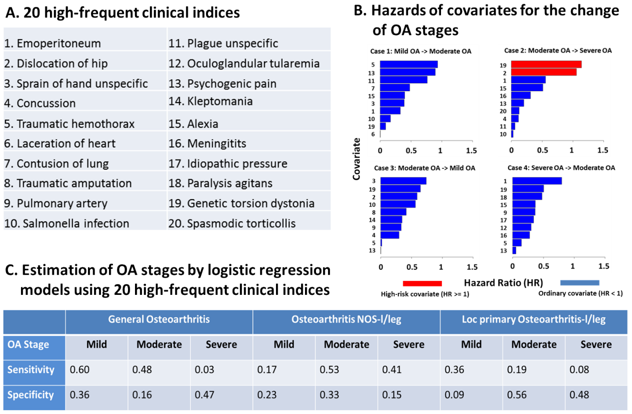
B. Hazards of covariates for the change of OA stages

Case 1: Mild OA -> Moderate OA
1 20 5 16 10 8 15 2 18 12
Case 2: Moderate OA -> Severe OA
4 5 18 10 11 14 19 3 20 13
Case 3: Moderate OA -> Mild OA
7 17 16 8 12 6 13 20 4 10
Case 4: Severe OA -> Moderate OA
7 10 3 1 13 18 15 19 16 8
Covariate
0 0.5 1 1.5
Hazard Ratio (HR)
High-risk covariate (HR >= 1)
Ordinary covariate (HR < 1)

C. Estimation of OA stages by logistic regression models using 20 high-frequent clinical indices

	General Osteoarthritis			Osteoarthritis NOS-l/leg			Loc primary Osteoarthritis-l/leg		
OA Stage	Mild	Moderate	Severe	Mild	Moderate	Severe	Mild	Moderate	Severe
Sensitivity	0.04	0.46	0.65	0.07	0.54	0.42	0.72	0.40	0.61
Specificity	0.32	0.44	0.38	0.50	0.37	0.39	0.37	0.49	0.53

FIGURE 8. Risk estimation of OA stages and comorbidities with partial correlation-based feature selection. (A). 20 high-frequent clinical indices generated by partial correlation; (B). Covariate hazards for OA stage progressions; (C). Estimation of OA stages by logistic regression models upon 20 high-frequent clinical indices.


A. 20 high-frequent clinical indices

1. Emoperitoneum	11. Plague unspecific
2. Dislocation of hip	12. Oculoglandular tularemia
3. Sprain of hand unspecific	13. Psychogenic pain
4. Concussion	14. Kleptomania
5. Traumatic hemothorax	15. Alexia
6. Laceration of heart	16. Meningitits
7. Contusion of lung	17. Idiopathic pressure
8. Traumatic amputation	18. Paralysis agitans
9. Pulmonary artery	19. Genetic torsion dystonia
10. Salmonella infection	20. Spasmodic torticollis

B. Hazards of covariates for the change of OA stages

Case 1: Mild OA -> Moderate OA
5 13 11 7 15 3 1 10 19 6
Case 2: Moderate OA -> Severe OA
19 2 1 15 16 13 20 4 11 10
Case 3: Moderate OA -> Mild OA
3 19 2 10 8 14 9 4 5 13
Case 4: Severe OA -> Moderate OA
1 19 18 15 9 17 12 16 5 13
Covariate
0 0.5 1 1.5
Hazard Ratio (HR)
High-risk covariate (HR >= 1)
Ordinary covariate (HR < 1)

C. Estimation of OA stages by logistic regression models using 20 high-frequent clinical indices

	General Osteoarthritis			Osteoarthritis NOS-l/leg			Loc primary Osteoarthritis-l/leg		
OA Stage	Mild	Moderate	Severe	Mild	Moderate	Severe	Mild	Moderate	Severe
Sensitivity	0.60	0.48	0.03	0.17	0.53	0.41	0.36	0.19	0.08
Specificity	0.36	0.16	0.47	0.23	0.33	0.15	0.09	0.56	0.48

FIGURE 9. Risk estimation of OA stages and comorbidities with adjusted cosine similarity-based feature selection. (A). 20 high-frequent clinical indices generated by adjusted cosine similarity; (B). Covariate hazards for OA stage progressions; (C). Estimation of OA stages by logistic regression models upon 20 high-frequent clinical indices.

Based on these clinical indices, the comorbidity risks and OA stages were estimated by Cox proportional hazard and logistic regression models. The panel B in Figure 7, 8 and 9 displays the impact of comorbidities on OA stage changes. We identified fewer clinical characteristics with hazards greater than 1 (red bars), which means that these comorbidities have an impact on OA progressions. For OA stage predictions, the panel C of Figure 7, 8 and 9 displays the sensitivity and specificity analysis results related to different stages for 3 types of OA. Regarding Table 4 (estimations based on features extracted by RDN), these feature selection methods give much lower statistical values which represent unreliable predictions.

We also delineated the association networks of clinical indices derived by the three feature selection methods.

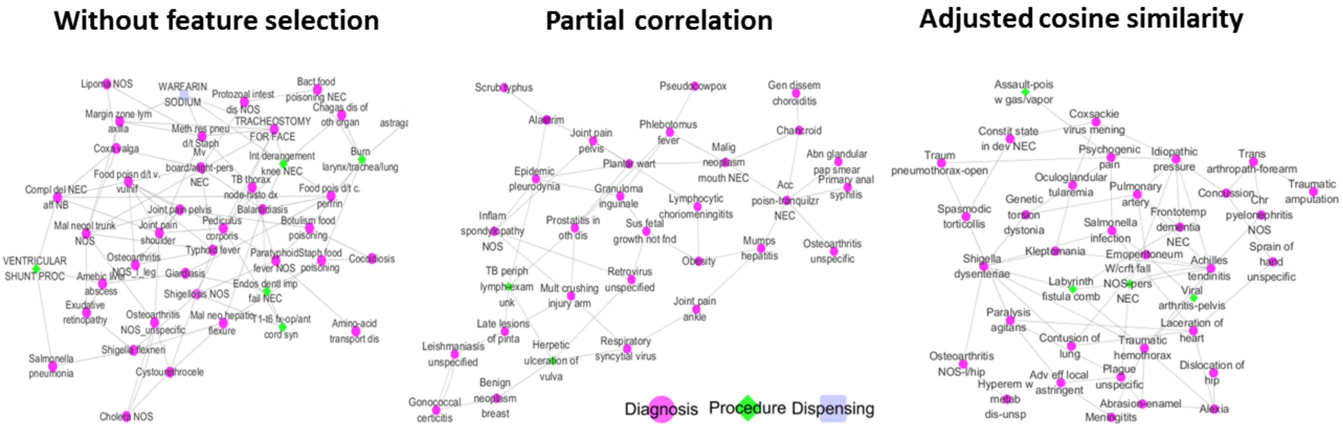


FIGURE 10. Association networks of clinical indices for the mild OA stage obtained without feature selection, partial correlation, and adjusted cosine similarity.

Figure 10 shows the results for the mild OA stage, which corresponds to Figure 2A (RDN for the mild stage of OA). In contrast, the size of RDN is larger than others. Meanwhile, the features in RDN refer to more clinical categories (diagnosis, procedure, DRG and dispensing). In addition, some associated indices in RDN were reported previously in the OA-related literature [5], [35], [36]. Therefore, RDN generates more reasonable clinical indices association networks.

In summary, our RDN-based pipeline can provide new and reasonable findings related to the impact of comorbidities on OA stage changes. RDNs extract representative co-existing diseases related to each of OA stages (Figure 2). Then, our new Cox proportional models derive core comorbidities affecting OA stage progressions (Figure 4). For the estimation of OA stages, our combined logistic models can provide reliable and reasonable predictions for specific stages (Table 4).

IV. DISCUSSION

Our approach pipeline is suitable for estimating comorbidities' impacts on the main disease with evident pathological stages. There exist numerous diseases with clear clinical phases such as Osteoarthritis, Chronic Kidney Disease, and Parkinson Disease. In addition, some co-existing diseases often play important roles in stage progressions (relief or aggravation). Through monitoring the status of comorbidities, the stages of the main disease can be estimated. Such practice is useful when it is inconvenient to measure the phases of the main disease. The accurate identification of disease stages will guide the medication or therapy decisions of related physicians (or clinical specialists).

The combined EMR-claims dataset provides a comprehensive data source for longitudinal tracing patients' clinical records in multiple healthcare institutions. Utilizing only TDW EMR data of one clinic may lead to the deficiency of complete diagnosis information, which brings tremendous missing values for important clinical factors. The estimation of OA staging, risk analyses and predictions of OA comorbidities will be unavailable or inaccurate. Using only claims data in different medical institutions will bring the loss of lab test results and some detailed symptom information. In addition, claims data does not include OA stage-related indexes (e.g. WOMAC). So the combination of TDW EMR with claims data can guarantee the integrity of medical records for specific patients during the censoring period. Based on the linked EMR-claims dataset, our current results are effective and reliable. Furthermore, we also tried to construct RDN by only using TDW EMR or claims dataset. The network size is evidently smaller than the present results. The corresponding clinical factors are incomplete. Therefore, the combination of TDW EMR and claims data can benefit the estimation of key comorbidities related to OA progression.

Our methodology can be summarized into 3 steps (Shown in Section "Introduction"). The 3 steps can be considered as a prototype to extract and analyze key comorbidities' impacts on the phase changes of the main disease. Through choosing specific statistical tests, BUFAM can flexibly handle the pairwise feature associations with various combinations among main data types (numeric, binary, categorical, ordinal). Cox proportional regression and GLM are combined to generate the prediction model of comorbidity risks and impacts on main disease progressions. In general, our prototype combines our proposed association analysis and classical risk estimation strategies. For the new patients' structural medical records, core comorbidities will be extracted accurately. Meanwhile, the related risks and effects can be analyzed exhaustively.

Referring to selecting association approaches for pairwise clinical features, we have tried the Pearson correlation previously. It shows that the Pearson correlation is more adaptive to numeric features. If the data types of two features are both numeric, Pearson correlation is a good choice. However, for other data types such as categorical, ordinal and binary, Pearson's performance is weaker than other statistical measurements (e.g. Spearman, one-way or chi-square test). Based on different combinations of data types, we have summarized the best association statistical tests in our BUFAM [17]. In our study, the association approaches are automatically selected by BUFAM based on the data types of pairwise clinical features.

The linked EMR-claims data density has impacts on RDNs. High-frequent clinical factors can help to generate more accurate RDNs. We selected 200 most frequently occurring factors for each patient on average. In this way, the appearance frequencies of specific clinical features are similar in most patients, which decreased the missing data's impacts on deriving RDNs.

For some cases, our pipeline is not suitable for the diseases with rapid progressions and low influences of comorbidities. Our framework is helpful to estimate comorbidities' impacts on the main disease with multiple co-existing symptoms and provide quantitative risk predictions related to the pathological stage progressions. Based on these estimations, physicians may make preventive or therapeutic plans for individual patients.

During the whole project, we have communicated with one of Rheumatology (Bursitis, Arthritis, Osteoarthritis) physicians in Wake Forest Baptist Medical Center. He confirmed some comorbidity impacts on the aggravation of OA symptoms such as Hypertension and Obesity. In addition, we have checked two literature related to comorbidities' frequency and impacts on pain and physical functions [37], [38]. Hypertension, dyslipidemia, obesity and ischaemic heart disease were reported as important comorbidities in general clinical practices, which match our partial results.

V. CONCLUSION

Our studies indicate that the combination of the EMR with claims data is a useful strategy for obtaining more accurate medical data sources from patients. The disease stage-dependent association networks, derived by RDNs, objectively reflect the detailed pathology of disease stages. Our analyses of the effects of clinical factors on OA staging clarify the associations between key covariates and OA progression. The prototype of risk estimations for comorbidities can be applied to any disease with evident pathological stages and comorbidity impacts.

DECLARATION OF INTEREST

None

CONTRIBUTORS

Prof. Xiaobo Zhou and Wei Chen conceived of the study design and total strategy. Wei Chen realized the work and wrote the manuscript. Kun Wei was in charge of the CDM format transformation and linked the EMR with claims data together. Dr. Weiling Zhao revised the manuscript and gave actionable guides on the whole work.

ACKNOWLEDGEMENT

The authors would like to thank the data extraction services of the Wake Forest Clinical and Translational Science Institute (WF CTSI), which is supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1TR001420. They would also like to thank the help of total colleagues in the Center of Systems Biology and Bioinformatics, Department of Diagnostic Radiology, Wake Forest Baptist Medical Center.

REFERENCES

[1] R. Caporali, M. A. Cimmino, P. Sarzi-Puttini, R. Scarpa, F. Parazzini, A. Zaninelli, A. Ciocci, and C. Montecucco, "Comorbid conditions in the AMICA study patients: Effects on the quality of life and drug prescriptions by general practitioners and specialists," *Seminars Arthritis Rheumatism*, vol. 35, pp. 31–37, Aug. 2005.

[2] Y. Zhang and J. M. Jordan, "Epidemiology of osteoarthritis," *Clinics Geriatric Med.*, vol. 26, pp. 355–369, Aug. 2010.

[3] W. H. Ettinger and R. F. Afable, "Physical disability from knee osteoarthritis: The role of exercise as an intervention," *Med. Sci. Sports Exerc.*, vol. 26, no. 12, pp. 1435–1440, 1994.

[4] A. R. Feinstein, "The pre-therapeutic classification of co-morbidity in chronic disease," *J. Chronic Diseases*, vol. 23, pp. 455–468, Dec. 1970.

[5] S. Zambon, P. Siviero, M. Denkinger, F. Limongi, M. V. Castell, S. van der Pas, Á. Otero, M. H. Edwards, R. Peter, N. L. Pedersen, M. Sánchez-Martinez, E. M. Dennison, A. Gesmundo, L. A. Schaap, D. J. H. Deeg, N. M. van Schoor, S. Maggi, and for the Eposa Research Group, "Role of osteoarthritis, comorbidity, and pain in determining functional limitations in older populations: European project on Osteoarthritis," *Arthritis Care Res.*, vol. 68, pp. 801–810, Jun. 2016.

[6] P. Siviero, S. Zambon, F. Limongi, M. V. Castell, C. Cooper, D. J. H. Deeg, M. D. Denkinger, E. M. Dennison, M. H. Edwards, A. Gesmundo, Á. Otero, N. L. Pedersen, R. Peter, R. Queipo, E. J. Timmermans, N. M. van Schoor, S. Maggi, and for the EPOSA Research Group, "How hand osteoarthritis, comorbidity, and pain interact to determine functional limitation in older people: Observations from the European project on OSteoArthritis study," *Osteoarthritis*, vol. 68, pp. 2662–2670, Nov. 2016.

[7] P. Suri, D. C. Morgenroth, and D. J. Hunter, "Epidemiology of osteoarthritis and associated comorbidities," *PM R*, vol. 4, pp. S10–S19, May 2012.

[8] M. Gore, K.-S. Tai, A. Sadosky, D. Leslie, and B. R. Stacey, "Clinical comorbidities, treatment patterns, and direct medical costs of patients with osteoarthritis in usual care: A retrospective claims database analysis," *J. Med. Econ.*, vol. 14, pp. 497–507, Jun. 2011.

[9] C. J. Lozada. (2015). *Progression of Osteoarthritis*. [Online]. Available: http://emedicine.medscape.com/article/1930582-overview

[10] R.-L. Hsieh, M.-T. Lo, W.-C. Liao, and W.-C. Lee, "Short-term effects of 890-nanometer radiation on pain, physical activity, and postural stability in patients with knee osteoarthritis: A double-blind, randomized, placebo-controlled study," *Arch. Phys. Med. Rehabil.*, vol. 93, pp. 757–764, May 2012.

[11] A. Raut and M. S. Gundeti, "Obesity and osteoarthritis comorbidity: Insights from Ayurveda," *J. Obesity Metabolic Res.*, vol. 1, pp. 89–94, Apr. 2014.

[12] M. Ruiz, S. Cosenza, M. Maumus, C. Jorgensen, and D. Noël, "Therapeutic application of mesenchymal stem cells in osteoarthritis," *Expert Opinion Biol. Therapy*, vol. 16, pp. 33–42, Sep. 2015.

[13] CMS. (2014). *Medicare Claims Processing Manual*. [Online]. Available: https://www.cms.gov/regulations-and-guidance/guidance/manuals/internet-only-manuals-ioms-items/cms018912.html

[14] J. Neville and D. Jensen, "Relational dependency networks," *J. Mach. Learn. Res.*, vol. 8, pp. 653–692, Mar. 2007.

[15] C. Combi, M. Gozzi, B. Oliboni, J. M. Juarez, and R. Marin, "Temporal similarity measures for querying clinical workflows," *Artif. Intell. Med.*, vol. 46, pp. 37–54, May 2009.

[16] C. E. Sluzki, "Personal social networks and health: Conceptual and clinical implications of their reciprocal impact," *Families, Syst., Health*, vol. 28, pp. 1–18, Mar. 2010.

[17] H. Chen, W. Chen, C. Liu, L. Zhang, J. Su, and X. Zhou, "Relational network for knowledge discovery through heterogeneous biomedical and clinical features," *Sci. Rep.*, vol. 6, Jul. 2016, Art. no. 29915.

[18] M. D. Natter, J. Quan, D. M. Ortiz, A. Bousvaros, N. T. Ilowite, C. J. Inman, K. Marsolo, A. J. McMurry, C. I. Sandborg, and L. E. Schanberg, "An i2b2-based, generalizable, open source, self-scaling chronic disease registry," *J. Amer. Med. Inform. Assoc.*, vol. 20, pp. 172–179, Jan. 2013.

[19] CMS. (2012). *Data Tables*. [Online]. Available: https://www.cms.gov/research-statistics-data-and-systems/research/mcbs/data-tables.html

[20] J. Goggins, K. Baker, and D. Felson, ''What WOMAC pain score should make a patient eligible for a trial in knee osteoarthritis?'' *J. Rheumatol.*, vol. 32, pp. 540–542, Mar. 2005.

[21] K. Fukunage and P. Narendra, ''A branch and bound algorithm for computing k-nearest neighbors,'' *IEEE Trans. Comput.*, vol. C-24, no. 7, pp. 750–753, Jul. 1975.

[22] F. Wolfe, ''Determinants of WOMAC function, pain and stiffness scores: Evidence for the role of low back pain, symptom counts, fatigue and depression in osteoarthritis, rheumatoid arthritis and fibromyalgia,'' *Rheumatology*, vol. 38, no. 4, pp. 355–361, Apr. 1999.

[23] H. C. Bhakta and C. A. Marco, ''Pain management: Association with patient satisfaction among emergency department patients,'' *J. Emergency Med.*, vol. 46, pp. 456–464, Apr. 2014.

[24] (2016). *Arthritis in Knee: 4 Stages of Osteoarthritis*. [Online]. Available: https://www.ibji.com/arthritis-in-knee-4-stages-of-osteoarthritis/

[25] J. M. Neuhaus, C. E. Mcculloch, and R. Boylan, ''Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes,'' *Statist. Med.*, vol. 32, pp. 2419–2429, Jun. 2013.

[26] W. Daniel and W. Wayne, ''Kruskal–Wallis one-way analysis of variance by ranks,'' *Appl. Nonparam. Statist.*, pp. 226–234, 1990.

[27] P. Verdecchia, F. Angeli, G. Mazzotta, P. Martire, M. Garofoli, G. Gentile, and G. Reboldi, ''Treatment strategies for osteoarthritis patients with pain and hypertension,'' *Therapeutic Adv. Musculoskeletal Disease*, vol. 2, pp. 229–240, Aug. 2010.

[28] L. K. King, L. March, and A. Anandacoomarasamy, ''Obesity & osteoarthritis,'' *Indian J. Med. Res.*, vol. 138, pp. 185–193, Aug. 2013.

[29] P. C. Austin, ''The performance of different propensity score methods for estimating marginal hazard ratios,'' *Statist. Med.*, vol. 32, pp. 2837–2849, Jul. 2013.

[30] M. Pavlou, G. Ambler, S. Seaman, O. Guttmann, P. Elliott, M. King, and R. Z. Omar, ''How to develop a more accurate risk prediction model when there are few events,'' *Res. Methods Reporting*, vol. 351, Aug. 2015, Art. no. h3868.

[31] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. Voelker, B. Nussenbaum, and E. W. Wang, ''A practical guide to understanding Kaplan-Meier curves,'' *Otolaryngol.–Head Neck Surg.*, vol. 143, pp. 331–336, Sep. 2010.

[32] R. Opgen-Rhein and K. Strimmer, ''Learning causal networks from systems biology time course data: An effective model selection procedure for the vector autoregressive process,'' *BMC Bioinf.*, vol. 8, p. S3, Mar. 2007.

[33] J. Yang, Y. Li, W. Cheng, Y. Liu, and C. Liu, ''EKF–GPR-based fingerprint renovation for subset-based indoor localization with adjusted cosine similarity,'' *Sensors*, vol. 18, p. 318, Jan. 2018.

[34] B. M. Sarwar. (2001). *Adjusted Cosine Similarity*. [Online]. Available: https://www10.org/cdrom/papers/519/node14.html

[35] S. Zambon, P. Siviero, M. Denkinger, F. Limongi, M. V. Castell, S. van der Pas, Á. Otero, M. H. Edwards, R. Peter, N. L. Pedersen, and M. Sánchez-Martinez, ''Osteoarthritis, comorbidity and pain: Their role in determining functional limitations in older populations (European project on Osteoarthritis),'' *Arthritis Care Res.*, vol. 68, no. 6, pp. 801–810, 2015.

[36] M. de Rooij, M. P. M. Steultjens, E. Avezaat, A. Häkkinen, A. Klaver, M. van der Leeden, T. Maas, L. D. Roorda, H. van der Velde, W. F. Lems, and J. Dekker, ''Restrictions and contraindications for exercise therapy in patients with hip and knee osteoarthritis and comorbidity,'' *Phys. Therapy Rev.*, vol. 18, pp. 101–111, Nov. 2013.

[37] A. A. Leite, A. J. G. Costa, B. de Arruda Matheos de Lima, A. V. L. Padilha, E. C. de Albuquerque, and C. D. L. Marques, ''Comorbidities in patients with osteoarthritis: Frequency and impact on pain and physical function,'' *Revista Brasileira de Reumatologia*, vol. 51, pp. 118–123, Apr. 2011.

[38] U. T. Kadam, K. Jordan, and P. R. Croft, ''Clinical comorbidity in patients with osteoarthritis: A case-control study of general practice consulters in England and Wales,'' *Ann. Rheumatic Diseases*, vol. 63, pp. 408–414, May 2003.