Journal of Pain Research

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Identification of a potential fibromyalgia diagnosis using random forest modeling applied to electronic medical records

This article was published in the following Dove Press journal:
Journal of Pain Research
10 June 2015
Number of times this article has been viewed

Birol Emir[1]
Elizabeth T Masters[1]
Jack Mardekian[1]
Andrew Clair[1]
Max Kuhn[2]
Stuart L Silverman[3]

[1]Pfizer Inc., New York, NY, [2]Pfizer Inc., Groton, CT, [3]Cedars-Sinai Medical Center, Los Angeles, CA, USA

Background: Diagnosis of fibromyalgia (FM), a chronic musculoskeletal condition characterized by widespread pain and a constellation of symptoms, remains challenging and is often delayed.

Methods: Random forest modeling of electronic medical records was used to identify variables that may facilitate earlier FM identification and diagnosis. Subjects aged ≥18 years with two or more listings of the International Classification of Diseases, Ninth Revision, (ICD-9) code for FM (ICD-9 729.1) ≥30 days apart during the 2012 calendar year were defined as cases among subjects associated with an integrated delivery network and who had one or more health care provider encounter in the Humedica database in calendar years 2011 and 2012. Controls were without the FM ICD-9 codes. Seventy-two demographic, clinical, and health care resource utilization variables were entered into a random forest model with downsampling to account for cohort imbalances (<1% subjects had FM). Importance of the top ten variables was ranked based on normalization to 100% for the variable with the largest loss in predicting performance by its omission from the model. Since random forest is a complex prediction method, a set of simple rules was derived to help understand what factors drive individual predictions.

Results: The ten variables identified by the model were: number of visits where laboratory/non-imaging diagnostic tests were ordered; number of outpatient visits excluding office visits; age; number of office visits; number of opioid prescriptions; number of medications prescribed; number of pain medications excluding opioids; number of medications administered/ordered; number of emergency room visits; and number of musculoskeletal conditions. A receiver operating characteristic curve confirmed the model's predictive accuracy using an independent test set (area under the curve, 0.810). To enhance interpretability, nine rules were developed that could be used with good predictive probability of an FM diagnosis and to identify no-FM subjects.

Conclusion: Random forest modeling may help to quantify the predictive probability of an FM diagnosis. Rules can be developed to simplify interpretability. Further validation of these models may facilitate earlier diagnosis and enhance management.

Keywords: fibromyalgia, random forest, predictive modeling, electronic medical records, health care resource utilization, real-world data

Background

Fibromyalgia (FM) is a chronic, complex musculoskeletal condition characterized by widespread pain generally defined as bilateral pain both above and below the waist and includes axial skeletal pain.[1,2] It has been well established that FM is associated with reduced patient function and quality of life as well as substantial health care resource utilization and associated costs.[3–7]

Correspondence: Birol Emir
Pfizer Inc., 235 East 42nd Street,
New York, NY 10017-5755, USA
Tel +1 212 733 8581
Fax +1 212 351 1008
Email birol.emir@pfizer.com

Diagnosis of FM has conventionally been based on the 1990 American College of Rheumatology classification criteria,[1] which was updated in 2010 by adding a symptom severity assessment, a widespread pain index, and eliminating the need for a tender point examination.[2] Despite these diagnostic criteria and the development of tools that may be useful to screen patients for the presence of FM,[8–10] diagnosis remains challenging, and patients tend to cycle through the health care system for years before being diagnosed with FM.[11–13]

The challenge of accurately diagnosing FM arises in part from the presence of a variety of symptoms in addition to pain, such as sleep disturbance, headache, and fatigue, as well as an association of FM with several comorbidities that include mood disorders, sleep disorders, and irritable bowel syndrome.[12,14,15] Thus, a search for specific characteristics or predictors of developing FM has been considered an important component of increasing the diagnostic accuracy and improving patient management. In the search for predictors, several studies have identified somatic symptoms, psychosocial and socioeconomic factors, fatigue, sleep problems, and workplace stress as significant precursors of widespread pain.[16–19] Another study that further explored predictors of FM identified several potential variables, including socioeconomic status, psychological distress, comorbidity, and rheumatoid arthritis severity.[20] However, that study only evaluated FM development in patients with rheumatoid arthritis, which may not necessarily reflect onset of FM in a broader population. The need to identify FM predictors was further emphasized in a recent narrative review of predictive FM studies.[21] The review discussed the association of FM with potential biological markers and clinical characteristics, but also highlighted the complexity of determining the importance of these variables as predictors, suggesting that additional studies or new approaches may be needed.

In addition to predictors of developing FM, another approach is to identify variables predictive of an FM diagnosis. Such an approach can inform health care providers of patients who may need specific evaluation for FM, which can facilitate earlier diagnosis and narrow the gap between symptom onset and diagnosis, thus also enhancing management strategies. To more accurately reflect the clinical setting, these variables are best identified using real-world data, in contrast to data from controlled clinical trials.

The availability of the electronic medical records (EMR) provides an opportunity to evaluate a wide array of variables associated with an FM diagnosis in the real-world clinical setting. Such records capture a variety of patient-level data that represent integral components of provider care that may not necessarily be available through other data sources such as administrative claims databases.[22] Predictive variables identified using EMR data may have greater applicability to clinical practice, and analyses of EMR data suggest that factors beyond demographic and clinical variables may be useful predictors of an FM diagnosis.[23,24] Our recent analysis of EMR data observed significant differences between FM and no-FM cohorts for most of the evaluated variables, including a greater prevalence of nearly all comorbidities and higher health care resource utilization across a range of resource categories.[24] The purpose of the current analysis was to use random forest modeling to expand on these differences, as univariate models do not account for relationships among variables, and to determine whether particular variables or sets of variables can be identified as predictive of an FM diagnosis. Random forest modeling is a computationally extensive data mining technique that can be used to identify and rank the importance of predictors from among the range of input variables. This technique uses historical data from subjects and attempts to accurately predict future outcomes from classification trees generated through data resampling to produce an integrated prediction that can be highly accurate. Random forest has previously been reported in the rheumatology setting for identifying genes predictive of rheumatoid arthritis[25] and factors predictive of knee arthroplasty in patients with osteoarthritis.[26] While it was also used to support the recent update of the FM diagnostic criteria,[2] the current study is the first to apply this technique to EMR data for predictive diagnostic purposes for FM. The predictive modeling approaches utilized in this study are consistent with the recently developed criteria for Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD).[27]

Materials and methods

Data source

Structured EMR data from the Humedica database, which longitudinally captures demographic, clinical, claims, and medical administrative information, were utilized for this analysis.[28] The Humedica database has broad geographic representation across the USA and aggregates deidentified EMR data from health care providers across the continuum of care including hospitals, medical groups, and integrated delivery networks. Patient records are linked using a unique patient identifier and are fully compliant with the Health Insurance Portability and Accountability Act with regard to

identification of patients and providers, as well as protected health information.

Subjects

All subjects who met the inclusion and exclusion criteria were included in the predictive modeling. Subjects identified for inclusion were those who were ≥18 years of age in 2011 and associated with an integrated delivery network with at least one encounter with a health care provider in the Humedica database in both 2011 and 2012. Exclusion criteria were the presence of at least one medical claim any time during 2011–2012 with an International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis code for malignant cancer (except for basal cell and squamous cell skin cancers and benign neoplasms); an ICD-9 code for diagnosis or procedure for transplantation; and residency in a nursing inpatient facility any time during 2011–2012. Subjects with these characteristics were excluded since they could confound the analysis due to high rates of resource use and a high prevalence of comorbid conditions relative to the populations of interest. The prediction model developed in this paper targets ambulatory patients with noncancer pain. Additionally, the presence of an FM diagnosis (ICD-9 code 729.1; myalgia and myositis, unspecified, which is the diagnostic code commonly used to identify FM) during subject enrollment prior to 2012 was also a reason for exclusion.

Among subjects who met all the inclusion and exclusion criteria, an FM cohort was defined as those subjects with at least two listings of the ICD-9 code 729.1 for FM at least 30 days apart during calendar year 2012, and the no-FM cohort consisted of similar subjects but without the ICD-9 codes for FM.

Predictive modeling

The total dataset was randomly divided into a training dataset (440,975 or 75% of subjects) to develop the model and a test dataset (146,985 or 25% of subjects) to confirm model performance. Splitting the data and allocating the test data to independently evaluate model performance attempts to eliminate the overfitting that can occur if relationships are identified in the training data that may not generally hold true.

Univariate analyses of potential predictors of FM were initially performed to explore differences in clinical and health care resource utilization variables between the FM and no-FM cohorts, and those results have been described elsewhere.[24] All 72 variables that were previously explored (see Table S1) were included in the current predictive modeling.

The objective of this analysis was to identify variables predictive of FM diagnosis by applying random forest predictive modeling to the EMR data. Random forest is a robust data mining technique with good predictive performance with respect to diagnostic accuracy.[29] Random forest models are ensembles of classification trees that are developed from a series of bootstrapped samples.[29] This technique was particularly attractive for the current analysis due to its relatively simple approach to handling severe imbalance in cohort sizes. In this dataset, the prevalence of FM was <1%, and therefore the sizes of the two groups of interest, FM and no-FM, were severely imbalanced; identifying cases in the prediction model when the outcome is rare is difficult for any prediction model. Therefore, we used an internal method within random forest called downsampling on the training dataset, which balances the number of FM and no-FM subjects at each bootstrap classification.[29] The test data were not adjusted for balance in order to provide more reliable estimates of the predictive performance of the model.

The final analysis on the training dataset incorporated the top ten predictor variables that were suggested by the random forest model, ranked by their importance (normalized to 100%) based on the variable with the largest loss in prediction performance by its omission in the model. A receiver operating characteristic (ROC) curve was generated using the test dataset to evaluate model performance. The random forest modeling was performed using R software (CRAN R) R 3.0 (cran.r-project.org/).

Enhancing interpretability

To enhance interpretability of the random forest models, cumulative distribution plots were developed for each of the predictive variables to illustrate determination of values for distinguishing between cohorts. These plots present the distribution of the cohorts across the range of values for each of the variables.

In addition to the cumulative distribution plots, a set of rules was developed using the C5.0 technique.[29] These rules generate sets of criteria that, when applied to subjects, identify subsets of subjects who have either a high predictive probability of FM or a high predictive probability of no-FM. To generate these rules, a simulated dataset was created in order to obtain a broader range of values for the ten predictors and to avoid concerns of overfitting through repeated use of the training dataset. The minimum, maximum, 20th, 40th, 60th, and 80th percentiles of the ten predictors identified by

the random forest model were computed using the training dataset. A total of 6^{10} possible combinations was created by using each combination of the ten predictors across the six percentiles. One example combination that was used consisted of all ten predictors at their minimum value. The simulated dataset was run through the random forest model to obtain a predicted probability of an FM diagnosis for each patient. Focusing on the simulated patients with the highest (≥0.70) and lowest (≤0.20) predicted probabilities of FM resulted in 4,179 simulated patients for analysis. A cutoff value of ≥0.70 was considered reasonable for classifying high predictive probability of FM and ≤0.20 was considered high predictive probability of no-FM, and the C5.0 rules were then applied to classify these patients. The rules identified thresholds among the predictive variables that were more likely to characterize FM and no-FM patients. These rules help to determine the patterns behind the predictive model and, as a group, help elucidate the reasons for each individual's prediction.

Results

Subject characteristics

As shown in Table 1, 587,961 subjects met all inclusion/exclusion criteria and had all required demographic and clinical information available in the Humedica database for this analysis during 2011 and 2012. Among these subjects, 4,296 (0.7%) were identified as having FM based on the predefined ICD-9 code criteria, resulting in 583,665 subjects in the no-FM cohort.

As shown in Table 2, significant differences were observed between the cohorts with regard to all demographic characteristics. The FM cohort was characterized by a higher predominance of females (78.7% versus 64.5%; $P<0.0001$) as well as differences in age, race, geographic distribution, and insurance plans. As previously described,[24] there were significant differences between the cohorts for clinical and health care resource utilization characteristics.

Table 1 Sample attrition table

Attrition criterion	n
Total number of deidentified patients in the Humedica database available for this research from January 1, 2011 to December 31, 2012	9,318,581
Patients aged ≥18 years in 2011	7,696,733
And enrolled in integrated delivery network	4,192,869
And with ≥1 encounter with a health care provider in 2011 and 2012	720,912
Patients with cancer diagnosis (exclusion)*	109,094
Patients with transplantation (exclusion)*	5,163
Patients with nursing home (exclusion)*	5,099
Patients with FM diagnosis prior to 2012 (exclusion)*	20,026
Excluding patients with cancer diagnosis, transplantation, in nursing home, or FM diagnosis prior to 2012	132,574
Patients meeting all inclusion and exclusion criteria	588,338
Missing sex value	377
Total patients in analysis population	587,961
≥2 ICD-9 codes for FM (729.1) at least 30 days apart during 2012	4,296
Number of patients in no-FM cohort	583,665

Notes: *Patients in these exclusion categories may have had more than one exclusion criterion.
Abbreviations: FM, fibromyalgia; ICD-9, International Classification of Diseases, Ninth Revision.

Random forest model

Figure 1 presents the top ten variables identified from the 72 variables input into the random forest model, and their relative importance to the model for predicting an FM diagnosis, normalized to 100% for the variable showing the greatest importance. Age was the only demographic variable identified in the top ten variables, and the number of musculoskeletal pain conditions was the only clinical variable; the other eight variables were a function of the magnitude of utilization of health care resource categories during 2011, the year prior

Table 2 Demographic characteristics of the evaluated cohorts

Variable	Value		*P*-value*
	FM (n=4,296)	No-FM (n=583,665)	
Female sex, n (%)	3,379 (78.7)	282,369 (64.5)	<0.0001
Age, years, mean (SD)	53.3 (14.6)	52.7 (16.3)	0.0318
Age distribution, n (%)			<0.0001
18–49 years	1,651 (38.4)	229,910 (39.4)	
50–64 years	1,482 (34.5)	183,414 (31.4)	
≥65 years	1,163 (27.1)	170,341 (29.2)	
Race, n (%)			<0.0001
African American	296 (6.9)	83,727 (14.3)	
Asian	32 (0.7)	11,294 (1.9)	
Caucasian	3,778 (87.9)	429,955 (73.7)	
Other/unknown	190 (4.4)	58,689 (10.1)	
Region, n (%)			<0.0001
Midwest	2,540 (59.1)	375,872 (64.4)	
Northeast	373 (8.7)	118,146 (20.2)	
South	1,125 (26.2)	75,414 (12.9)	
West	5 (0.1)	458 (0.1)	
Other/unknown	253 (5.9)	13,775 (2.4)	
Insurance type, n (%)			<0.0001
Commercial	181 (4.2)	145,425 (24.9)	
Medicaid	7 (0.2)	3,740 (0.6)	
Medicare	88 (2.0)	61,151 (10.5)	
Missing/unknown	4,017 (93.5)	370,332 (63.4)	
Other payer type	0	297 (0.1)	
Uninsured	3 (0.1)	2,720 (0.5)	
Charlson Comorbidity Index, mean (SD)	0.8 (1.3)	0.5 (1.1)	<0.0001

Note: *Age and Charlson Comorbidity Index means were compared using two-sample *t*-tests and categorical variables were compared using chi-square tests.
Abbreviations: FM, fibromyalgia; SD, standard deviation.

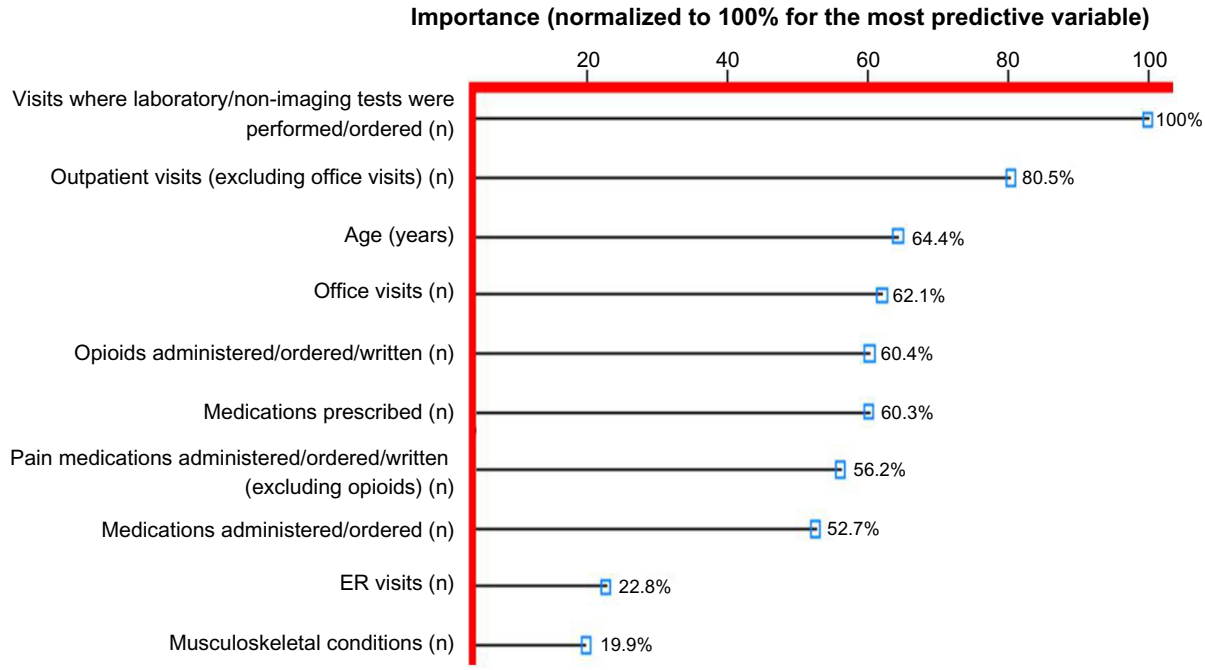

Figure 1 The ten most important variables for predicting a diagnosis of fibromyalgia identified from random forest models.
Notes: The level of importance, as shown on the x-axis, ranked for all identified variables based on normalization to 100% for the variable with the largest loss in predicting performance by its omission in the model.
Abbreviation: ER, emergency room.

to the FM diagnosis. The most important predictive variable was the "number of visits during which diagnostic/laboratory tests were ordered", followed by the "number of outpatient visits excluding office visits", which had an importance of 80.5%, and "age" ranked third (64.4%). There was a cluster of variables in the range of 50%–60% importance, most of which were related to medication utilization, followed by "number of ER visits" and "number of musculoskeletal conditions", both of which appeared to have substantially lower importance, 22.8% and 19.9%, respectively.

A receiver operating characteristic (ROC) curve was generated to evaluate the sensitivity and specificity of the predicted probabilities versus observed outcome when the model was run using the test dataset. The ROC curve shown in Figure 2 had an area under the curve (c-statistic) of 0.810, indicating good accuracy for predicting an FM diagnosis. The ROC curve also shows that at a cut-off probability of 0.500, sensitivity was 0.641 and specificity was 0.794, and that the optimal balance of sensitivity (0.721) and specificity (0.740) results in an estimated cutoff probability value of 0.446 (Figure 2).

Enhancing model interpretability

Cumulative distribution plots were developed to show the range of values for each of the predictor variables and to display the differences between FM and no-FM subjects. Figure 3A shows 70% of cases had ≤3 visits where laboratory/diagnostic testing was ordered compared with approximately 90% of no-FM subjects. Similarly, as shown in Figure 3E, 36% of FM subjects had more than two opioid prescriptions ordered compared with approximately 10%

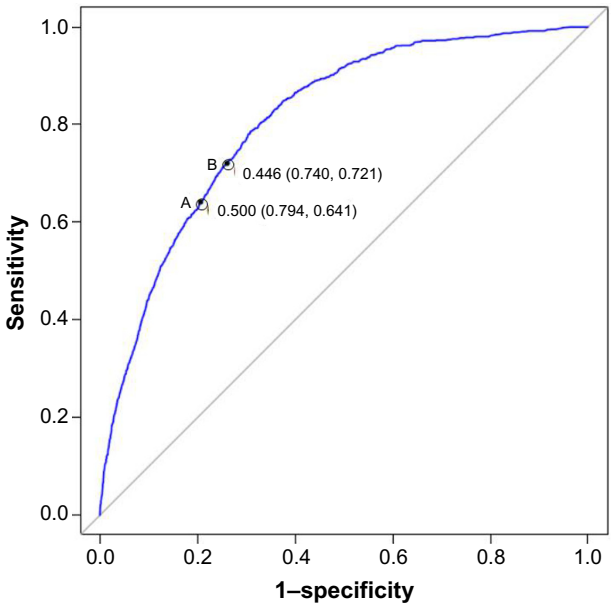

Figure 2 Receiver operating characteristic curve modeled using the test dataset.
Notes: Receiver operating characteristic curve of the sensitivity and specificity for predicting the probability of a fibromyalgia diagnosis modeled using the test dataset from the ten most important variables identified from the random forest model. Point A, which denotes a probability value of 0.500, has a sensitivity of 0.641 and a specificity of 0.794. In contrast, point B shows the probability value, 0.446, that provides balance between sensitivity (0.721) and specificity (0.740).

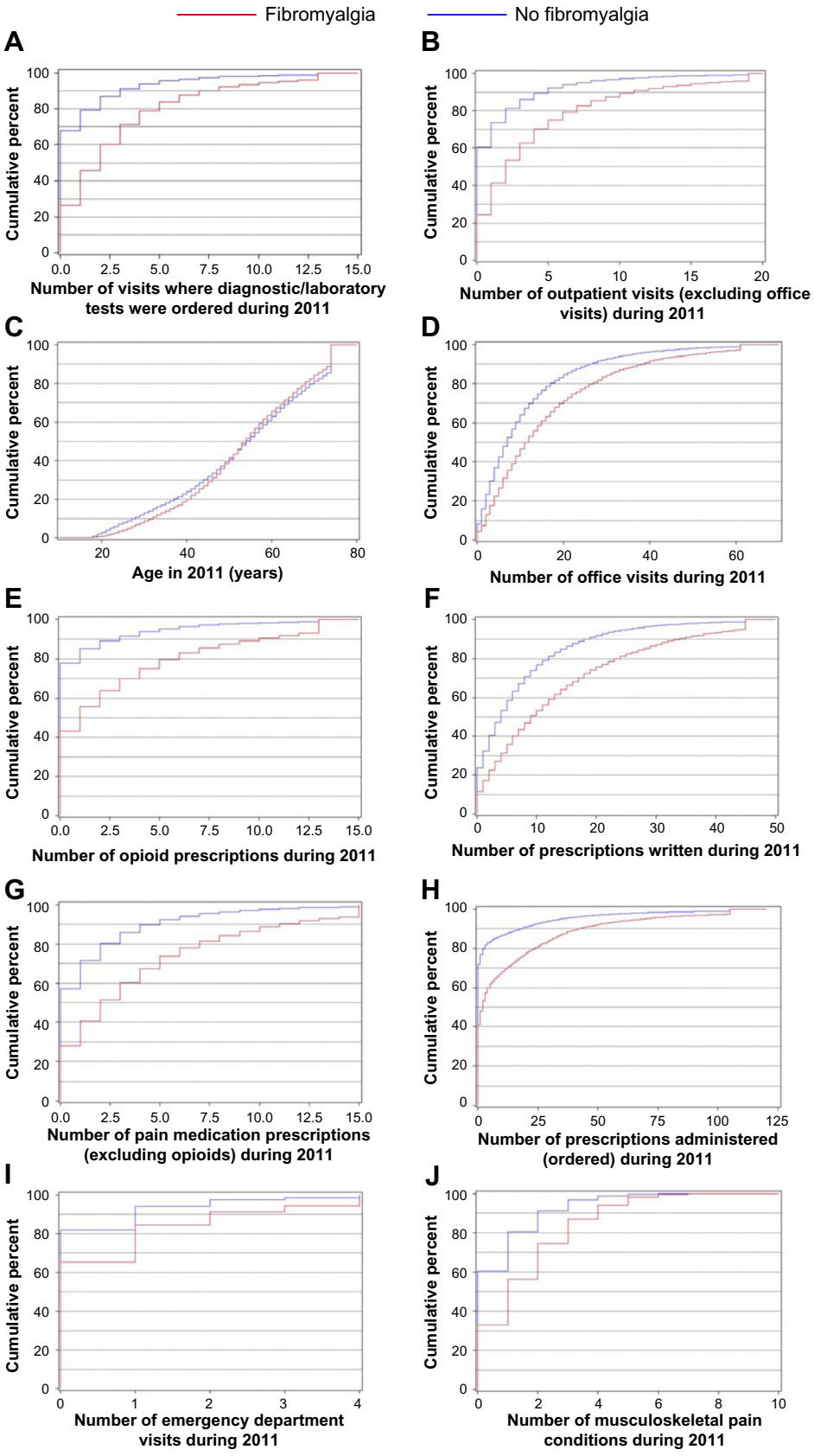

Figure 3 Cumulative distribution functions for the variables identified in the random forest model.

Notes: (**A**) Number of visits during which diagnostic/laboratory tests were ordered. (**B**) Number of outpatient visits (excluding office visits). (**C**) Age. (**D**) Number of office visits. (**E**) Number of opioid prescriptions. (**F**) Number of prescriptions written. (**G**) Number of pain medication prescriptions (excluding opioids). (**H**) Number of prescriptions administered (ordered). (**I**) Number of emergency department visits. (**J**) Number of musculoskeletal pain conditions.

Table 3 Rules for identifying FM and no-FM subjects based on results of the predictive modeling using a technique known as C5.0 rules

Rule number	Predictive class	Rule (all components must be met)	Number of subjects predicted in simulated dataset (n=4,179) to belong to predictive class	Percentage of subjects in simulated dataset (n=4179) correctly identified in predictive class	Sensitivity (%) computed in patients identified by rule applied to test dataset (n=146,985)	Specificity (%) computed in patients identified by rule applied to test dataset (n=146,985)
1	FM	Number of outpatient visits >0 Number of prescriptions administered ≤3 Number of musculoskeletal pain conditions >0	308	99.7	78.3	39.7
2	FM	Number of visits where laboratory/non-imaging tests were ordered >0 Number of musculoskeletal pain conditions >0	247	99.6	85.6	26.6
3	FM	Number of outpatient visits >0 Number of office visits ≤9 Number of opioid prescriptions >0	208	99.5	75.9	34.9
4	FM	Number of visits where laboratory/non-imaging tests were ordered >0 Number of emergency room visits >0	102	99	94.8	15.4
5	FM	Number of visits where laboratory/non-imaging tests were ordered >0 Number of pain medications excluding opioids >2	63	98.5	92.7	18.5
6	No-FM	Number of visits where laboratory/non-imaging tests were ordered =0 Number of opioid prescriptions =0 Number of musculoskeletal pain conditions =0	2,176	100.0	99.6	0
7	No-FM	Number of opioid prescriptions =0 Number of pain medications excluding opioids ≤2 Number of emergency room visits =0 Number of musculoskeletal pain conditions =0	1,761	99.9	96.6	5.6
8	No-FM	Number of visits where laboratory/non-imaging tests were ordered =0 Number of office visits >9	1,224	99.9	94.7	36.3
9	No-FM	Number of visits where laboratory/non-imaging tests were ordered =0 Number of outpatient visits =0	3,091	99	98.2	15.8

Abbreviation: FM, fibromyalgia.

of no-FM subjects. The largest difference between cohorts can be seen in Figure 3A (number of visits where laboratory tests and/or non-imaging diagnostic tests were ordered), 3E (number of opioid medications), 3F (number of medications), and 3G (number of pain medications excluding opioids).

In addition, a rule-based approach enabled development of nine sets of rules (Table 3), any one of which could be used to determine whether a subject is likely to be diagnosed with FM as long as each component of the rule is satisfied. As an example, of 4,179 predictor cases in the prediction dataset, 308 cases satisfied the conditions of rule 1, and 99.7% of these cases (307 of 308) were correctly identified with predictor values associated with a high predicted FM probability (ie, ≥0.70, which was considered reasonable as a high cutoff value for classifying a patient as having an FM diagnosis, Table 3). The implication is that a subject with characteristics satisfied by rule 1 has a high potential for an FM diagnosis. Similarly, rule 6 selected 2,176 cases, all with predictor values leading to a low predicted FM probability (≤0.20), indicating a high potential to be a no-FM subject. For each of the rules, sensitivity was high based on the test dataset (75.9%–99.6%), but specificity was low (0%–39.7%, Table 3).

Discussion

This analysis is the first to apply random forest methodology to EMR data for the purpose of predictive modeling of a musculoskeletal diagnosis. It expands on a recent univariate analysis that reported significant differences between FM and no-FM cohorts across a range of demographic, clinical, and health care resource utilization variables extracted from EMR data.[24] While that analysis showed which variables were associated with an FM diagnosis, the current analysis evaluated how these variables perform as predictors of an FM diagnosis. The results show that eight of the ten most important variables identified as being predictive of an FM diagnosis were related to health care resource utilization. Only age among the demographic characteristics and only number of musculoskeletal pain conditions among the clinical characteristics were included in the top ten predictors.

A preponderance of health care resource utilization variables as predictors of an FM diagnosis is not entirely surprising, given the high rate of health care resource utilization that has consistently been observed in FM populations.[30–32] The relevance and importance of these variables as predictors of an FM diagnosis are further supported by the observations that there is high resource use even in the years before a definitive FM diagnosis,[11,33] likely resulting from the patient's journey in search for an explanation of their symptoms.[11] In particular, the two most important predictive variables from the random forest model were "number of visits where laboratory/non-imaging tests were ordered" and "number of outpatient visits", with rankings of 100% and 80% importance, respectively.

Interestingly, although FM is more prevalent in women, sex was not a top predictive variable, and age was the only demographic predictor identified in the model. These results may be due to the fact that the variables were evaluated for predicting an FM diagnosis, rather than the characteristics predictive of the disease.

Good model performance was supported by ROC analysis, with a c-statistic of 0.810, within the range of 0.8–0.9 considered as having good accuracy for a diagnostic test. More practically, sets of rules were developed to differentially evaluate the likelihood of FM or no-FM diagnoses. Multiple rules enable a broad choice for determining the likelihood of FM based on the availability of data for the specific predictive variables. In the current analysis, application of C5.0 rules attempted to derive simple rules to elucidate which factors drive individual predictions. Given the severe imbalance between FM and no-FM cases in the test dataset, the low specificity resulting from applying the rules to identify FM cases and the low specificity resulting from applying the rules to identify no-FM cases is not surprising.

This analysis complements other studies that have evaluated biologic markers for predicting FM development (reviewed by Ablin and Buskila[21]). While biologic markers are important for understanding the etiology and pathogenesis of FM, diagnostic predictors may have greater direct clinical application for making evaluation and treatment decisions, with the goal of reducing the patient and economic burdens associated with FM. Additionally, although various predictive models and algorithms are available that can be applied to administrative databases, random forest may be especially appropriate for use in FM for several reasons. These reasons include the need to account for the low prevalence of FM in the database and the established good predictive properties of this technique. It should also be noted that random forest has previously been applied to FM as described in the updated American College of Rheumatology diagnostic criteria.[2] However, that application used random forest to determine the symptoms of greatest importance that physicians use to recognize FM. In contrast, the current analysis was not restricted to symptoms, but used a wide array of variables available from EMR to not only identify predictors

of an FM diagnosis, but also define sets of these variables that can be applied to enhance predictive probability in the clinical setting.

Interpretation and generalizability of these results should consider both the strengths and limitations of the study. The strength of this study is its external validity resulting from use of "real-world" EMR data comprised of elements captured in routine clinical practice from multiple sites. EMR contain patient-level data, including many types of data that are not generally available in claims databases, which enable tracking of individuals longitudinally. Since such datasets have not previously been applied to predictive FM modeling, this also represents a new and novel approach for evaluating FM. However, the data source also represents a limitation, since as with all such database studies, there is the potential for errors in coding or record-keeping at the point of the health care provider. In this regard, these analyses were predicated on the validity of an FM diagnosis, which also represents a limitation, especially since the criteria used by providers to diagnose FM are not uniformly collected. In order to improve the accuracy of identifying such subjects, the presence of two or more ICD-9 codes for FM was required for inclusion. Further support for the use of this method may be obtained in a validation study by examining individual charts to verify the accuracy of the diagnosis based on ICD-9 coding, and such a study may be warranted.

Another limitation is that the variables that were identified are those associated with an FM diagnosis rather than characteristics associated with the disease itself. However, it may also be considered that the source of the data and the models used provide a foundation for identifying EMR markers for diagnosis of a disease that is as yet incompletely characterized with regard to readily recognized biomarkers. At the least, the use of predictive modeling described here would identify a subset of individuals who may require more comprehensive screening for FM based on high health care resource utilization. The observational nature of this study is also a limitation, since causal inferences cannot be made and all results should be considered inferential.

In summary, random forest modeling can be applied to determine the likelihood of an FM diagnosis. Use of cumulative distribution plots or development of predictive rules for application in the clinical setting can simplify this method with good accuracy. These types of analyses go beyond questionnaires that are available for patient screening and can be directly applied to a variety of clinical variables that are available through EMR. The variables identified in these analyses help to describe characteristics of patients ultimately receiving an FM diagnosis as identified through EMR, thereby providing clinicians with additional information to aid in the understanding of this condition. The value of this approach is in identifying patients who may require more comprehensive screening for FM, thereby also potentially reducing the delay in diagnosis and treatment that often occurs. Further validation of random forest models may enhance diagnostic and management strategies for FM.

Acknowledgment

Editorial assistance was provided by E Jay Bienen, who was funded by Pfizer Inc.

Disclosure

This research was sponsored by Pfizer Inc. ETM, JM, BE, AC, and MK are employees and shareholders of Pfizer Inc., the sponsor of this study. SLS was not financially compensated for his collaboration on this project or for the development of this manuscript.

References

1. Wolfe F, Smythe HA, Yunus MB, et al. The American College of Rheumatology 1990 Criteria for the Classification of Fibromyalgia. Report of the Multicenter Criteria Committee. *Arthritis Rheum*. 1990; 33(2):160–172.
2. Wolfe F, Clauw DJ, Fitzcharles MA, et al. The American College of Rheumatology preliminary diagnostic criteria for fibromyalgia and measurement of symptom severity. *Arthritis Care Res*. 2010;62(5): 600–610.
3. Hoffman DL, Dukes E. The health status burden of people with fibromyalgia: a review of studies that assessed health status with the SF-36 or the SF-12. *Int J Clin Pract*. 2008;62(1):115–126.
4. Salaffi F, Sarzi-Puttini P, Girolimetti R, Atzeni F, Gasparini S, Grassi W. Health-related quality of life in fibromyalgia patients: a comparison with rheumatoid arthritis patients and the general population using the SF-36 health survey. *Clin Exp Rheumatol*. 2009;27(5 Suppl 56):S67–S74.
5. Wolfe F, Michaud K, Li T, Katz RS. EQ-5D and SF-36 quality of life measures in systemic lupus erythematosus: comparisons with rheumatoid arthritis, noninflammatory rheumatic disorders, and fibromyalgia. *J Rheumatol*. 2010;37(2):296–304.
6. Luo X, Cappelleri JC, Chandran A. The burden of fibromyalgia: assessment of health status using the EuroQol (EQ-5D) in patients with fibromyalgia relative to other chronic conditions. *Health Outcomes Res Med*. 2011;2(4):e203–e214.
7. Schaefer C, Chandran A, Hufstader M, et al. The comparative burden of mild, moderate and severe fibromyalgia: results from a cross-sectional survey in the United States. *Health Qual Life Outcomes*. 2011; 9(1):71.
8. White KP, Harth M, Speechley M, Ostbye T. Testing an instrument to screen for fibromyalgia syndrome in general population studies: the London Fibromyalgia Epidemiology Study Screening Questionnaire. *J Rheumatol*. 1999;26(4):880–884.
9. Perrot S, Bouhassira D, Fermanian J. Development and validation of the Fibromyalgia Rapid Screening Tool (FiRST). *Pain*. 2010;150(2): 250–256.
10. Arnold LM, Stanford SB, Welge JA, Crofford LJ. Development and testing of the fibromyalgia diagnostic screen for primary care. *J Womens Health (Larchmt)*. 2012;21(2):231–239.

11. Choy E, Perrot S, Leon T, et al. A patient survey of the impact of fibromyalgia and the journey to diagnosis. *BMC Health Serv Res*. 2010;10:102.
12. Arnold LM, Clauw DJ, McCarberg BH. Improving the recognition and diagnosis of fibromyalgia. *Mayo Clin Proc*. 2011;86(5):457–464.
13. Amital H, Bar-On Y, Shalev V, Weitzman D, Chodick G. Understanding the factors influencing time to diagnosis in fibromyalgia. *Arthritis Rheumatol*. 2014;66 Suppl 11:S907.
14. Bennett RM, Jones J, Turk DC, Matallana L. An Internet survey of 2,596 people with fibromyalgia. *BMC Musculoskelet Disord*. 2007;8:27.
15. Gore M, Tai K-S, Chandran A, Zlateva G, Leslie D. Clinical comorbidities, treatment patterns, and healthcare costs among patients with fibromyalgia newly prescribed pregabalin or duloxetine in usual care. *J Med Econ*. 2012;15(1):19–31.
16. McBeth J, MacFarlane GJ, Benjamin S, Silman AJ. Features of somatization predict the onset of chronic widespread pain: results of a large population-based study. *Arthritis Rheum*. 2001;44(4):940–946.
17. McBeth J, MacFarlane GJ, Hunt IM, Silman AJ. Risk factors for persistent chronic widespread pain: a community-based study. *Rheumatology (Oxford)*. 2001;40(1):95–101.
18. MacFarlane GJ, Norrie G, Atherton K, Power C, Jones GT. The influence of socioeconomic status on the reporting of regional and widespread musculoskeletal pain: results from the 1958 British Birth Cohort Study. *Ann Rheum Dis*. 2009;68(10):1591–1595.
19. Gupta A, Silman AJ, Ray D, et al. The role of psychosocial factors in predicting the onset of chronic widespread pain: results from a prospective population-based study. *Rheumatology (Oxford)*. 2007;46(4):666–671.
20. Wolfe F, Hauser W, Hassett AL, Katz RS, Walitt BT. The development of fibromyalgia – I: examination of rates and predictors in patients with rheumatoid arthritis (RA). *Pain*. 2011;152(2):291–299.
21. Ablin JN, Buskila D. Predicting fibromyalgia, a narrative review: are we better than fools and children? *Eur J Pain*. 2014;18(8):1060–1066.
22. Hayrinen K, Saranto K, Nykanen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform*. 2008;77(5):291–304.
23. Masters ET, Mardekian J, Clair A, Silverman S. Identifying predictors of a fibromyalgia diagnosis: a retrospective electronic medical record analysis. *Arthritis Rheum*. 2013;65 Suppl:S52.
24. Masters ET, Mardekian J, Emir B, Clair A, Kuhn M, Silverman S. Electronic medical record data to identify variables associated with a fibromyalgia diagnosis: importance of healthcare resource utilization. *J Pain Res*. 2015;8:131–138.
25. Tang R, Sinnwell JP, Li J, Rider DN, de Andrade M, Biernacka JM. Identification of genes and haplotypes that predict rheumatoid arthritis using random forests. *BMC Proc*. 2009;3 Suppl 7:S68.
26. Riddle DL, Kong X, Jiranek WA. Two-year incidence and predictors of future knee arthroplasty in persons with symptomatic knee osteoarthritis: preliminary analysis of longitudinal data from the osteoarthritis initiative. *Knee*. 2009;16(6):494–500.
27. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1–W73.
28. Research and analytics powered by clinical data. Boston, MA, USA: Humedica, Inc.; 2013. Available from: http://www.humedica.com/solutions/research/. Accessed January 23, 2015.
29. Kuhn M, Johnson K. *Applied Predictive Modeling*. New York, NY, USA: Springer; 2013.
30. White LA, Birnbaum HG, Kaltenboeck A, Tang J, Mallett D, Robinson RL. Employees with fibromyalgia: medical comorbidity, healthcare costs, and work loss. *J Occup Environ Med*. 2008;50(1):13–24.
31. Lachaine J, Beauchemin C, Landry PA. Clinical and economic characteristics of patients with fibromyalgia syndrome. *Clin J Pain*. 2010;26(4):284–290.
32. Knight T, Schaefer C, Chandran A, Zlateva G, Winkelmann A, Perrot S. Health-resource use and costs associated with fibromyalgia in France, Germany, and the United States. *Clinicoecon Outcomes Res*. 2013;5:171–180.
33. Hughes G, Martinez C, Myon E, Taieb C, Wessely S. The impact of a diagnosis of fibromyalgia on health care resource use by primary care patients in the UK. An observational study based on clinical practice. *Arthritis Rheum*. 2006;54(1):177–183.

Supplementary materials

Table S1 Variables put into random forest model

Demographic variables
Age
Sex
Race
Clinical variables
Charlson comorbidity: myocardial infarction
Charlson comorbidity: congestive heart failure
Charlson comorbidity: peripheral vascular disease
Charlson comorbidity: cerebrovascular disease
Charlson comorbidity: dementia
Charlson comorbidity: chronic pulmonary disease
Charlson comorbidity: rheumatologic disease
Charlson comorbidity: peptic ulcer disease
Charlson comorbidity: mild liver disease
Charlson comorbidity: diabetes
Charlson comorbidity: diabetes with chronic complications
Charlson comorbidity: hemiplegia or paraplegia
Charlson comorbidity: renal disease
Charlson comorbidity: moderate or severe liver disease
Atypical facial pain
Autonomic neuropathies
Anxiety/generalized anxiety disorder
Back and neck pain (other than lower back pain)
Back and neck pain with neuropathic involvement (excluding low back)
Bipolar disorder
Causalgias
Chest pain
Arthritis and other arthropathies
Carpal tunnel syndrome
Myocardial infarction/congestive heart failure/peripheral vascular disease/cerebrovascular disease/coronary heart disease/hypertension/hyperlipidemia
Interstitial cystitis
Diffuse diseases of connective tissue
Depression
Dyspareunia
Chronic fatigue syndrome
Gastroesophageal reflux disease/gastritis/duodenitis/other gastrointestinal disease
Headache/migraine
Irritable bowel syndrome
Insomnia/sleep disorders/sleep apnea
Low back pain
Restless leg syndrome
Lupus
Memory loss
Mononeuritis of lower limb
Neuritis radiculitis
Osteoarthritis
Other musculoskeletal pain conditions
Other polyneuropathies
Panic disorder
Phantom limb pain
Postherpetic neuralgia
Post-traumatic stress disorder
Rheumatoid arthritis
Rheumatism (excluding the back)
Tinnitus
Temporomandibular joint disorder
Trigeminal neuralgia
Number of musculoskeletal pain conditions
Number of neuropathic pain conditions
Diagnosis of obesity
Health care resource utilization variables
Acupuncture
Chiropractic visit
Counseling (exercise counseling, nutrition counseling)
Number of emergency department visits
Number of visits where imaging was ordered
Number of hospitalizations
Number of visits where diagnostic/laboratoy tests were ordered
Number of office visits
Number of other outpatient visits
Physical therapy
Number of total prescriptions administered (ordered)
Number of total prescriptions written
Number of opioid prescriptions
Number of total pain medication prescriptions (excluding opioids)