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

Birol Emir1
Elizabeth T Masters1
Jack Mardekian1
Andrew Clair1
Max Kuhn2
Stuart L Silverman3
1Pfizer Inc., New York, NY, 2Pfizer Inc., Groton, CT, 3Cedars-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
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.\(^12\)\(^–\)\(^15\) 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 $\geq 18$ 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. 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. Random forest models are ensembles of classification trees that are developed from a series of bootstrapped samples. 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. 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. 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 (\(\geq 0.70\)) and lowest (\(\leq 0.20\)) predicted probabilities of FM resulted in 4,179 simulated patients for analysis. A cutoff value of \(\geq 0.70\) was considered reasonable for classifying high predictive probability of FM and \(\leq 0.20\) 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, there were significant differences between the cohorts for clinical and health care resource utilization characteristics.

### 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>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>(P)-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female sex, (n) (%)&lt;br&gt;FM ((n=4,296))&lt;br&gt;No-FM ((n=583,665))</td>
<td>3,379 (78.7)&lt;br&gt;282,369 (64.5)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Age, years, mean (SD)</td>
<td>53.3 (14.6)&lt;br&gt;52.7 (16.3)</td>
<td>0.0318</td>
</tr>
<tr>
<td>Age distribution, (n) (%)&lt;br&gt;18–49 years</td>
<td>1,651 (38.4)&lt;br&gt;1,703 (34.6)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>50–64 years</td>
<td>1,482 (34.5)&lt;br&gt;1,834 (36.4)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>(\geq 65) years</td>
<td>1,125 (26.2)&lt;br&gt;20,026 (41.7)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Race, (n) (%)&lt;br&gt;African American</td>
<td>296 (6.9)&lt;br&gt;83,727 (17.4)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Asian</td>
<td>32 (0.7)&lt;br&gt;11,294 (2.4)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Caucasian</td>
<td>3,778 (87.9)&lt;br&gt;429,955 (87.9)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Other/unknown</td>
<td>190 (4.4)&lt;br&gt;58,689 (12.2)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Region, (n) (%)&lt;br&gt;Midwest</td>
<td>2,540 (59.1)&lt;br&gt;375,872 (80.4)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Northeast</td>
<td>375 (8.7)&lt;br&gt;118,146 (25.2)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>South</td>
<td>1,125 (26.2)&lt;br&gt;75,414 (15.9)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>West</td>
<td>5 (0.1)&lt;br&gt;458 (0.1)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Other/unknown</td>
<td>253 (5.9)&lt;br&gt;13,775 (2.9)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Insurance type, (n) (%)&lt;br&gt;Commercial</td>
<td>181 (4.2)&lt;br&gt;145,425 (29.4)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Medicaid</td>
<td>7 (0.2)&lt;br&gt;3,740 (0.8)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Medicare</td>
<td>88 (2.0)&lt;br&gt;61,151 (13.0)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Missing/unknown</td>
<td>4,017 (93.5)&lt;br&gt;370,332 (83.4)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Other payer type</td>
<td>0&lt;br&gt;297 (0.1)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Uninsured</td>
<td>3 (0.1)&lt;br&gt;2,720 (0.5)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Charlson Comorbidity Index, mean (SD)</td>
<td>0.8 (1.3)&lt;br&gt;0.5 (1.1)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

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; ICD-9, International Classification of Diseases, Ninth Revision.
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 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

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

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. 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. 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, likely resulting from the patient’s journey in search for an explanation of their symptoms. 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). 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. 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

### 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
- Causalgia
- 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
- Mononeuropathy of lower limb
- Neuritis radiculitis
- Osteoarthritis
- Other musculoskeletal pain conditions
- Other polyneuropathies
- Panic disorder
- Phantom limb pain
- Postherpetic neuralgia
- Post-traumatic stress disorder

#### 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/laboratory 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)