Arthritis & Rheumatology
Vol. 71, No. 12, December 2019, pp 1987–1996
DOI 10.1002/art.41056

AMERICAN COLLEGE of RHEUMATOLOGY
Empowering Rheumatology Professionals

Machine Learning to Predict Anti–Tumor Necrosis Factor Drug Responses of Rheumatoid Arthritis Patients by Integrating Clinical and Genetic Markers

Yuanfang Guan,[1] Hongjiu Zhang,[1] Daniel Quang,[1] Ziyan Wang,[1] Stephen C. J. Parker,[1] Dimitrios A. Pappas,[2] Joel M. Kremer,[3] and Fan Zhu[4]

Objective. Accurate prediction of treatment responses in rheumatoid arthritis (RA) patients can provide valuable information on effective drug selection. Anti–tumor necrosis factor (anti-TNF) drugs are an important second-line treatment after methotrexate, the classic first-line treatment for RA. However, patient heterogeneity hinders identification of predictive biomarkers and accurate modeling of anti-TNF drug responses. This study was undertaken to investigate the usefulness of machine learning to assist in developing predictive models for treatment response.

Methods. Using data on patient demographics, baseline disease assessment, treatment, and single-nucleotide polymorphism (SNP) array from the Dialogue on Reverse Engineering Assessment and Methods (DREAM): Rheumatoid Arthritis Responder Challenge, we created a Gaussian process regression model to predict changes in the Disease Activity Score in 28 joints (DAS28) for the patients and to classify them into either the responder or the nonresponder group. This model was developed and cross-validated using data from 1,892 RA patients. It was evaluated using an independent data set from 680 patients. We examined the effectiveness of the similarity modeling and the contribution of individual features.

Results. In the cross-validation tests, our method predicted changes in DAS28 (ΔDAS28), with a correlation coefficient of 0.405. It correctly classified responses from 78% of patients. In the independent test, this method achieved a Pearson's correlation coefficient of 0.393 in predicting ΔDAS28. Gaussian process regression effectively remapped the feature space and identified subpopulations that do not respond well to anti-TNF treatments. Genetic SNP biomarkers showed small contributions in the prediction when added to the clinical models. This was the best-performing model in the DREAM Challenge.

Conclusion. The model described here shows promise in guiding treatment decisions in clinical practice, based primarily on clinical profiles with additional genetic information.

INTRODUCTION

Rheumatoid arthritis (RA) patients show great heterogeneity in their responses to treatments (1), and accurate prediction of these responses would provide valuable information for optimal drug selection (2). In current practice, patients who respond inadequately to conventional therapies usually receive anti–tumor necrosis factor (anti-TNF) drugs as a second-line therapy (3). These expensive drugs are mainly chosen on a trial-and-error basis (4), and ~30% of patients respond poorly to them (5). Nonresponders incur costly drug expenses (6) and experience unimproved disease conditions (6,7), treatment side effects (8,9), and infection risks (10). In order to provide effective treatment to these patients, physicians need the ability to predict in advance how individual patients will respond to various anti-TNF drugs.

Supported by the National Science Foundation (1452656) and National Natural Science Foundation of China (61802360).

[1]Yuanfang Guan, PhD, Hongjiu Zhang, PhD (current address: Microsoft, Seattle, Washington), Daniel Quang, Ziyan Wang, Stephen C. J. Parker, PhD: University of Michigan, Ann Arbor; [2]Dimitrios A. Pappas, MD: Columbia University College of Physicians and Surgeons, New York, New York, and Corrona LLC, Waltham, Massachusetts; [3]Joel M. Kremer, MD: Corrona LLC, Waltham, Massachusetts, and Albany Medical College and the Center for Rheumatology, Albany, New York; [4]Fan Zhu, PhD: Chinese Academy of Sciences, Chongqing, China.

Drs. Guan, Zhang, and Zhu contributed equally to this work.

Dr. Guan has received consulting fees, speaking fees, and/or honoraria from Eli Lilly, Genentech, Hoffmann-La Roche, and Cleerly Inc. (more than $10,000 each), owns equity shares at Cleerly Inc. and Ann Arbor Algorithms Inc., and receives research support from Merck KGaA and Ryss Tech. Dr. Zhang owns stock in Microsoft. No other disclosures relevant to this article were reported.

Address correspondence to Yuanfang Guan, PhD, 500 South State Street, Ann Arbor, MI 48109. E-mail: gyuanfan@umich.edu.

Submitted for publication May 21, 2018; accepted in revised form July 18, 2019.

However, patient heterogeneity hinders the identification of predictive biomarkers and accurate modeling of anti-TNF responses. Early studies showed that demographic and clinical markers such as sex and baseline disease activity are related to treatment responses (11,12), but these markers are not predictive enough independently to identify nonresponders (13). Recent genome-wide association studies have identified multiple genetic markers that are associated with poor drug responses, and these variants were used to facilitate the identification of nonresponders (5,14–16). However, results obtained using these variant markers are often confounded by cohort or ethnic group (17). RA patients exhibit great genotypic and phenotypic heterogeneity, and markers found in one ethnic population or cohort may not be applicable to others (18). Thus, effective modeling of patient heterogeneity is the key to accurate drug response prediction.

In the present study, we used a Gaussian process regression (GPR) model to predict anti-TNF responses. This model was awarded first-place in the Dialogue on Reverse Engineering Assessment and Methods (DREAM): Rheumatoid Arthritis Responder Challenge (19,20). The GPR model combines demographic, clinical, and genetic markers, predicts changes in disease activity scores 24 months after baseline assessment, and identifies nonresponders to anti-TNF treatments. Specifically, the model predicts changes in the Disease Activity Score in 28 joints (DAS28) (21) of patients who have received 12 months of anti-TNF treatment and also classifies patient responses according to the European League Against Rheumatism response criteria (22). Using this model, we examined the transformation that the GPR kernel applied to the patient data as well as the distribution of the patients in the transformed space. Using an independent testing data set, we aimed to evaluate the potential of the GPR model in improving anti-TNF drug selection and the effect of genetic markers across multiple cohorts.

PATIENTS AND METHODS

Data acquisition. The training and testing data sets used in this study were provided by the DREAM Challenge organizers (Table 1). Data on 1,892 of these 2,706 individuals (chosen randomly) were provided to the participants in the competition before the final evaluation. Data on the remaining subjects were withheld for real-time submission evaluation, so that participants could assess their models throughout the competition. None of these data were included in the final evaluation. The samples in the training data set consisted of patients of European ancestry from the following 13 cohorts: the Autoimmune Biomarkers Collaborative Network (US) (23), the Genetics Network Rheumatology Amsterdam (The Netherlands), the Behandelstrategieen voor Rheumatoide Arthritis (The Netherlands) (24), the Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate (UK) (25,26), the Brigham Rheumatoid Arthritis Sequential Study (US) (27), the Epidemiological Investigation of Rheumatoid Arthritis (Sweden) (28), the Immunex Early Rheumatoid Arthritis study (US) (29), the Karolinska Institutet study (Sweden), the collection from Leiden University Medical Center (The Netherlands), the Treatment of Early Aggressive Rheumatoid Arthritis study (US), the Dutch Rheumatoid Arthritis Monitoring registry, the ApotheekZorg database (The Netherlands) (30,31), and the Research in Active Rheumatoid Arthritis trial (France) (32). All patients were either diagnosed as having RA by a board-certified rheumatologist or met the 1987 American College of Rheumatology criteria for RA (33). All patients had a baseline DAS28 of >3.2.

Table 1. Demographic, treatment, and response information on the training and testing (CERTAIN cohort) data sets*

	Training data set (n = 1,892)	CERTAIN cohort data set (n = 680)
Demographic data		
Mean age, years	54.9	55.6
Female, %	75.1	77.9
Treatment		
Methotrexate	1,332 (70.4)	441 (64.9)
Adalimumab	757 (40.0)	210 (30.9)
Etanercept	520 (27.5)	179 (26.3)
Infliximab	609 (32.2)	177 (26.0)
Certolizumab	0 (0)	114 (16.8)
Baseline DAS28, mean	5.87	4.73
Response status		
Nonresponder	436 (23)	238 (35)
ΔDAS28, mean	2.15	1.17

* Except where indicated otherwise, values are the number (%) of patients. CERTAIN = Comparative Effectiveness Registry to study Therapies for Arthritis and Inflammatory Conditions; DAS28 = Disease Activity Score in 28 joints.

The testing data were collected from 680 patients in the Corrona registry (34) who participated in the Comparative Effectiveness Registry to study Therapies for Arthritis and Inflammatory Conditions (CERTAIN) study (35). The CERTAIN study was conducted using data from the Corrona registry and involved adult RA patients who were diagnosed by certified rheumatologists, who had at least moderate disease activity defined by a Clinical Disease Activity Index (36) of >10, and who were starting or switching biologic agents.

For all patients in both the training and the testing data sets, information on sex, age, methotrexate use, and baseline DAS28 were collected and provided. Posttreatment DAS28 scores of all patients in the training data sets were available to all participants in the DREAM Challenge, whereas those of patients in the testing data sets were withheld by the organizers until the end of the challenge. For each subject, a panel of genotype imputation was provided (https://www.synapse.org/#!Synapse:syn1734172/wiki/62201).

Treatment response prediction. The proposed model in this study adopted GPR to predict changes in the DAS28 (ΔDAS28). GPR is designed to predict the unknown dependent variable for any given independent variable(s) based on known but noisy observations of the dependent and independent variables.

GPR does not match its target function to a specific model (e.g., linear, quadratic, or cubic). In this study, the GPR model took the input ΔDAS28 of known subjects as noisy observations and predicted treatment responses based on patients' clinical and genetic features. For more information on the full formulation, see Supplementary Table 1, available at on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.41056/abstract. The final model is available at https://www.synapse.org/#!Synapse:syn2368045/wiki/64596.

The kernel function of the GPR model accepts the differences in demographic data, treatment, and genetic features between 2 patients as input variables. For each input variable, the kernel function performs a squared exponential transformation and takes the summation of the transformed values. The distance between each pair of patients is then simultaneously determined by the nonlinear transformed difference across all features.

The genetic features included in the model were chosen via literature search or statistical analysis (Supplementary Table 2, http://onlinelibrary.wiley.com/doi/10.1002/art.41056/abstract). Genetic features were collected from the literature based on their reported association with either RA risks or anti-TNF responses; those from statistical analysis were collected based on the correlation between the genotype dosages and the ΔDAS28. During the DREAM Challenge, the organizing team provided a leaderboard set (used to compare the performance of all participants) to test the models. Both sets of candidate features were then tested using cross-validation and forward feature selection, and only features that improved the cross-validation performance and were on the leaderboard were included. Any single-nucleotide polymorphism (SNP) collected from the literature that did not show substantial correlation with the drug response phenotype was nevertheless retained if it added to the prediction performance. In the end, our model included 10–15 SNPs from the literature (depending on drug type) and ~5 SNPs from correlation testing. In addition, SNPs were tested using FaST-LMM (37), a linear mixed model–based program for performing both single-SNP and SNP-set genome-wide significance testing. However, none of the genetic features passed the 5×10^{-8} threshold in the FaST-LMM test when clinical variables were available.

All clinical and demographic features provided in the data sets were included in the model. Separate models were developed for different anti-TNF drugs. Each model was trained using data only from patients who received the corresponding anti-TNF treatment. In order to predict the responses of patients receiving certolizumab, a general model was developed using all patients in the training data set, but without genetic features. Models to classify patients as responders or nonresponders were developed using a binary cutoff of 0 (responders) or 1 (nonresponders) based on the predicted values, so that the proportion of nonresponders was consistent with that in the training set.

Evaluation. The prediction models were evaluated using 5-round 2-fold cross-validation (38). Each round of the repeated tests started with randomly splitting the training data set into 2 halves. A model was trained on 1 half and scored on the other, and the same was done with both halves swapped. Five rounds of tests yielded a total of 10 scores, and their average was used as the estimated performance score of the model. Repeated cross-validation tests were performed on the training data set provided by the DREAM Challenge. Because we were investigating 3 different drugs, we carried out evaluations on each of them as well as on the combined set.

The predicted ΔDAS28 was evaluated using Pearson's correlation coefficient. The performance of nonresponder identification was measured using the correctly predicted percentage and area under the receiver operating characteristic curve (AUC). The receiver operating characteristic curve was created by plotting the true-positive ratio (the ratio of the number of correctly identified patients to the number of all subjects with Alzheimer's disease or mild cognitive impairment) against false-positive ratio (the ratio of the number of incorrectly identified patients to the number of all subjects) at various threshold settings.

In order to assess the improvement of the GPR models compared to other algorithms, models using different algorithms were assessed via bootstrap test. This test is similar to the test method used in the original DREAM Challenge to assess the robustness of the models. These tests were repeated 100 times. In each of these rounds of testing, the training data set was the same size as the original challenge training data set (n = 1,892) and was sampled from the original challenge training data with replacement. The unsampled data were merged as the testing data. Performance was evaluated by examining the number of bootstrap tests in which GPR outperformed other methods.

RESULTS

Accuracy of GPR predicting anti-TNF responses. Given data on patient demographics, baseline DAS28, treatment, and SNP array, the GPR model predicted ΔDAS28 for patients and classified them as responders or nonresponders (Figure 1). It relied on a custom kernel function to weight a subject proportionally according to his or her similarity to the paired patient based on clinical and genetic data. The model predicted treatment outcomes by leveraging the outcomes and features of training patients, and it estimated patients' ΔDAS28 24 months after the initial disease assessment and classified them as responders or nonresponders. It would be expected that the estimations derived with our method would be close to the outcomes in patients with similar demographic, treatment, and SNP array data. We developed separate GPR models for different anti-TNF treatments.

There was a major difficulty in developing this method due to heterogeneity in the data sets. The DREAM Challenge used different cohorts for training and testing, requiring the participating

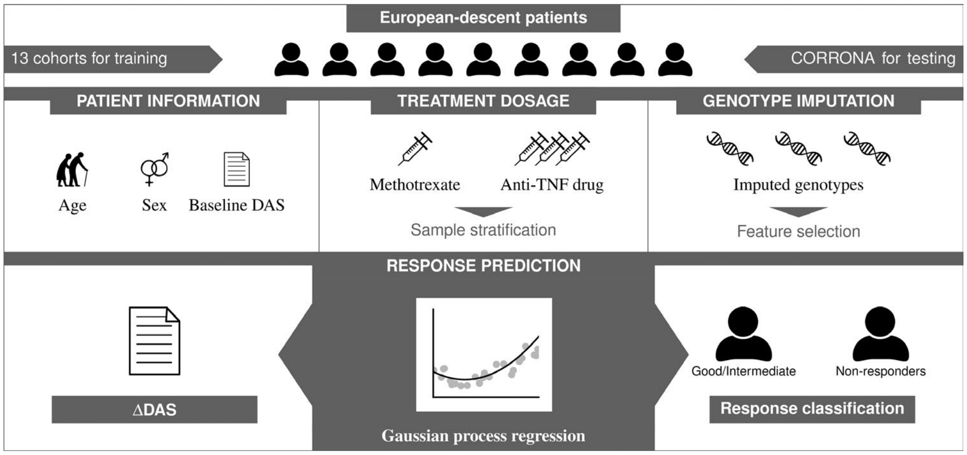

Figure 1. Overview of the treatment response prediction model. The model uses data on patient demographics, baseline Disease Activity Score in 28 joints (DAS28), treatment, and single-nucleotide polymorphism array to predict ΔDAS28 for the patient and to classify the patient as a responder or nonresponder. Anti-TNF = anti–tumor necrosis factor.

models to take into account the cohort effects. Post-challenge analysis on the 2 data sets, using the Mann-Whitney U test, showed that DAS28 ($P < 10^{-23}$) and ΔDAS28 ($P < 10^{-17}$) were drastically different. To address this heterogeneity, the GPR model was able to use data on similar patients from the training data set to guide the prediction, without explicitly specifying the feature distribution across cohorts. We then compared the results obtained with our GPR model to those obtained using other alternative models.

We evaluated the GPR model through repeated cross-validation tests as described above. The GPR model was compared to linear models, classification and regression tree models, and a support vector machine (SVM) model. For ΔDAS28 prediction, GPR achieved the best average correlation (0.405) between predicted and observed ΔDAS28, followed by ridge regression, SVM, and regression tree models (Figure 2B). In terms of response classification, the GPR model was shown to be the best performer overall, correctly classifying ~78% of subjects, with an AUC of ~0.66 (Figures 2C and D). Another well-performing model, SVM, is also a kernel-based method that shares properties with GPR. These methods can be used in clinical settings to inform the decision-making process. Compared to random assignment (50/50 chance of classification accuracy), using a GPR or SVM model increased the rate of accurate classification, and therefore of accurate treatment selection, by 28%. To evaluate the margin of improvement from our GPR models, we performed 100 bootstrap tests on the original data set and

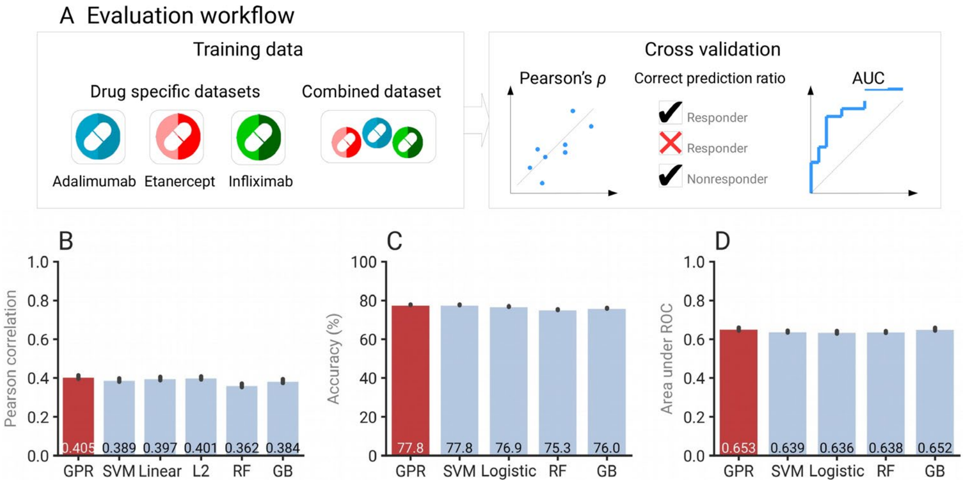

Figure 2. **A**, Overview of repeated cross-validation evaluation. All models underwent both treatment-specific and overall evaluations and were measured based on the 3 listed metrics. **B**, Pearson's correlation coefficients between the observed change in Disease Activity Score in 28 joints and predictions from tested regression methods. **C**, Accuracy (percentage correct classification) of tested responder versus nonresponder classification methods. **D**, Area under the receiver operating characteristic (ROC) curve (AUC) of tested responder versus nonresponder classification methods. Values are the mean ± SEM. The final model is displayed in red. GPR = Gaussian process regression; SVM = support vector machine; L2 = L2 regularization; RF = random forest; GB = gradient boosting regression/decision tree.

found that GPR showed substantial improvement over all other methods (Supplementary Table 1, http://onlinelibrary.wiley.com/doi/10.1002/art.41056/abstract). We also evaluated the models for individual drugs separately (Supplementary Figure 1, http://onlinelibrary.wiley.com/doi/10.1002/art.41056/abstract). The performance rankings varied, but the GPR model achieved the best performance overall, closely followed by linear and SVM models. We further optimized the hyperparameter in the model (Supplementary Figure 2, http://onlinelibrary.wiley.com/doi/10.1002/art.41056/abstract). The GPR model was then submitted for the DREAM Challenge evaluation.

Evaluation on an independent "hidden" cohort. The DREAM Challenge also evaluated the GPR model using an independent testing cohort (20). The independent data set, released after the competition, consisted of information on 680 patients from the CERTAIN study, conducted by Corrona (35). The GPR model achieved a Pearson's correlation coefficient of 0.393 when predicting ΔDAS28 (*P* < 1^{e-6} versus null hypothesis) and an AUC of 0.615 when classifying anti-TNF nonresponders. This represents reduction in correlation coefficient of only ~0.01 from the cross-validation, indicating limited overfitting and batch effects. The GPR model showed more consistent prediction performance than both the cross-validation data set and the independent testing data set.

Treatment response prediction enhanced by similarity modeling. To investigate how GPR effectively modeled patient heterogeneity, we inspected the properties of individual features. Among all the features, baseline DAS28 had the highest correlation with ΔDAS28 (Supplementary Table 3, http://onlinelibrary.wiley.com/doi/10.1002/art.41056/abstract). Across all of the samples in the training data set, regardless of specific anti-TNF treatment, the correlation coefficient between the baseline DAS28 and the ΔDAS28 was 0.370. In the CERTAIN cohort, the correlation coefficient was 0.351. However, the high correlation between baseline DAS28 and ΔDAS28 does not fully explain the performance of the GPR model. The difference in the performance of the GPR model and that of a naive baseline DAS28 linear regression implies some contribution of other demographic, clinical, and genetic features.

GPR relies on its kernel function to transform input features. To study how the kernel transformation incorporates features to help predict treatment responses, we projected the training sample in the feature spaces before and after the kernel transformation (Figure 3 and Supplementary Figures 3 and 4, http://onlinelibrary.wiley.com/doi/10.1002/art.41056/abstract). Principal components analysis on the features showed major confounding factors such as geographic or cohort information. The major contributing factors to the first 2 principal components were genetic features, which categorized patients based on their cohort information instead of their treatment responses. Conversely, principal components analysis

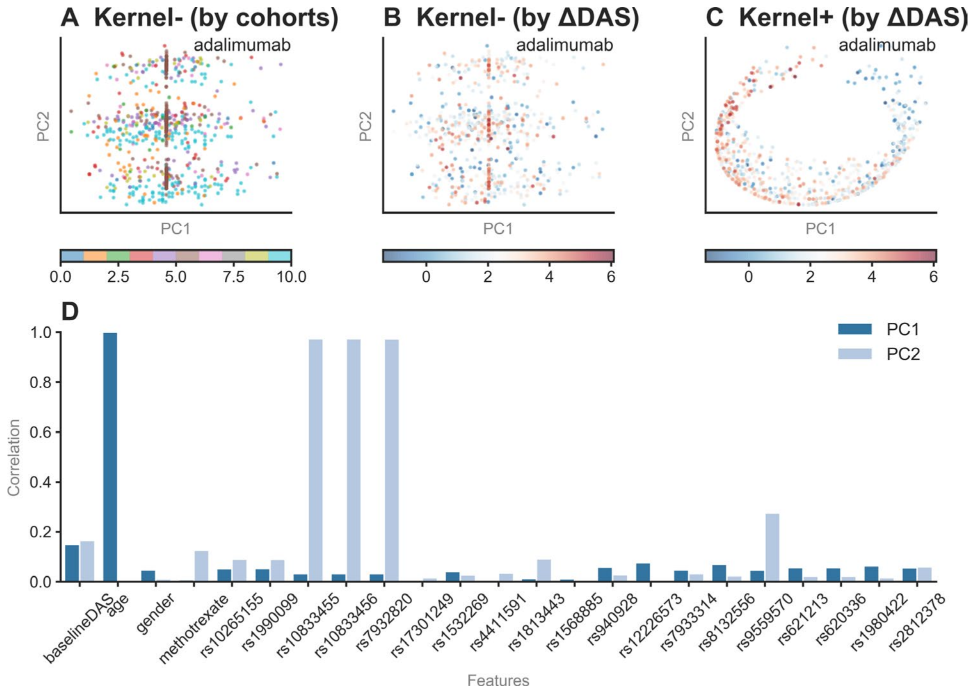

Figure 3. Feature space analysis of adalimumab users in the training data set. **A**, Principal components analysis (PCA) of the original feature space (without kernel transformation, colored according to change in Disease Activity Score in 28 joints [ΔDAS28]) shows separation by cohort. **B**, PCA of the original feature space (without kernel transformation, colored according to cohort) does not show obvious separation of responders and nonresponders. **C**, PCA of the kernel matrix (colored according to ΔDAS28) shows a clear gradient from responders to nonresponders. **D**, Feature contributions to first 2 principal components, based on the findings shown in **C**.

on the kernel-transformed similarity matrix showed a clear gradient from anti-TNF responders to nonresponders. Important features besides baseline DAS28, such as age, methotrexate use, and several genetic markers, correlated well with the first 2 principal components. The pattern demonstrated that in the kernel-transformed feature space, patients' similarity correlated well with their similarity in ΔDAS28 rather than other confounding factors.

Clinical and genetic heterogeneity in cross-cohort evaluation. During the assessment of features described above, we noticed a large variation in features among patients receiving different anti-TNF treatments. Therefore, in the original submission to the DREAM Challenge, we developed specialized models for individual drugs. In the post-challenge analysis, we reassessed the treatment-specific approach on the testing cohort. Our original analysis using the training data set showed that the baseline DAS28 correlated more strongly with ΔDAS28 among patients taking adalimumab and etanercept than those taking infliximab. However, in the CERTAIN cohort, baseline DAS28 scores were substantially less predictive of responses to adalimumab and infliximab than responses to etanercept. To accommodate the variation, we excluded patients receiving different drugs from kernel calculation for each drug-specific model. We also curated drug-specific genetic feature lists. We compared these models to two sets of genetic feature–free models, one trained on all patients and the other on those receiving only one corresponding anti-TNF treatment. The performance of these approaches was similar (Supplementary Figure 5, http://onlinelibrary.wiley.com/doi/10.1002/art.41056/abstract). The negligible difference produced by sample exclusion suggests that variation across different drugs may be attributed to sampling biases instead of treatment-specific characteristics. The conflict between the aforementioned high contributions of genetic features in treatment-specific models and high performance of clinical information–only models suggested the need for further analysis on the roles of genetic markers in treatment response prediction.

We therefore next investigated the contribution of genetic markers to treatment response prediction. The genetic markers we investigated were chosen based on either literature review or statistical analysis on the training data set (Supplementary

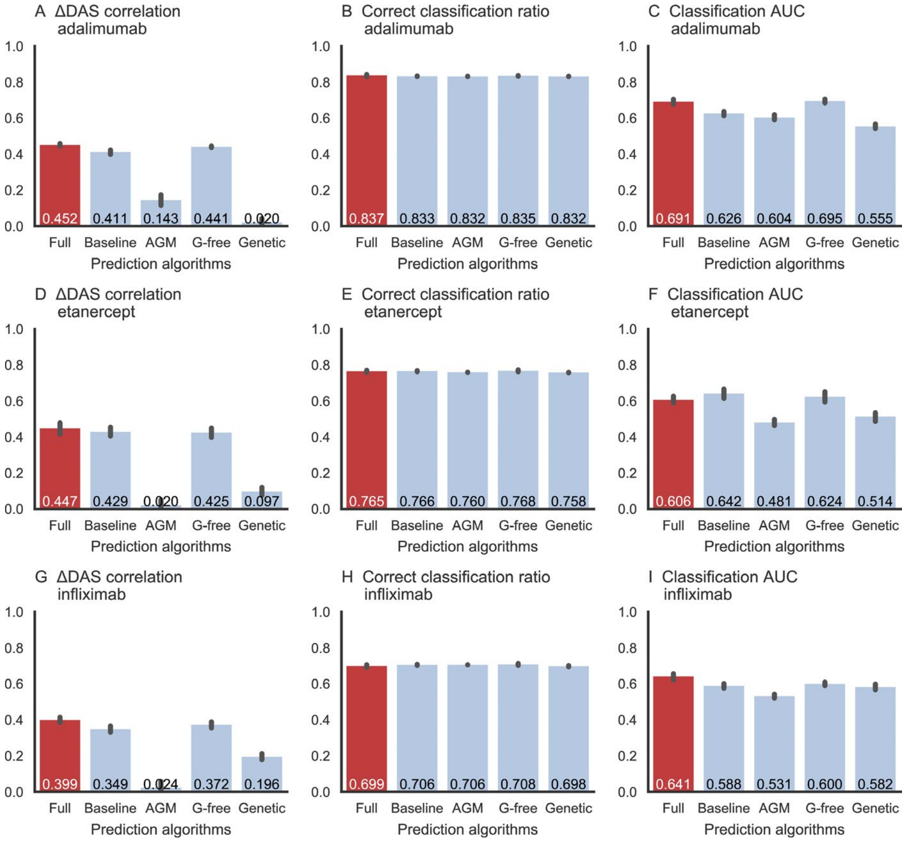

Figure 4. Repeated cross-validation tests of models with different feature sets using the training data set. **A**, **D**, and **G**, Pearson's correlation coefficients between the observed change in Disease Activity Score in 28 joints (ΔDAS28) and predictions from tested regression methods for adalimumab, etanercept, and infliximab. **B**, **E**, and **H**, Classification accuracy ratio of predictions from tested responder versus nonresponder classification methods for adalimumab, etanercept, and infliximab. **C**, **F**, and **I**, Area under the receiver operating characteristic curve (AUC) of tested responder versus nonresponder classification methods for adalimumab, etanercept, and infliximab. Model variations were developed to include all features, baseline DAS28 score only, age/sex (G)/methotrexate use (AGM) only, features except for sex (G-free), and genetic features only. Values are the mean SEM. The final model is displayed in red.

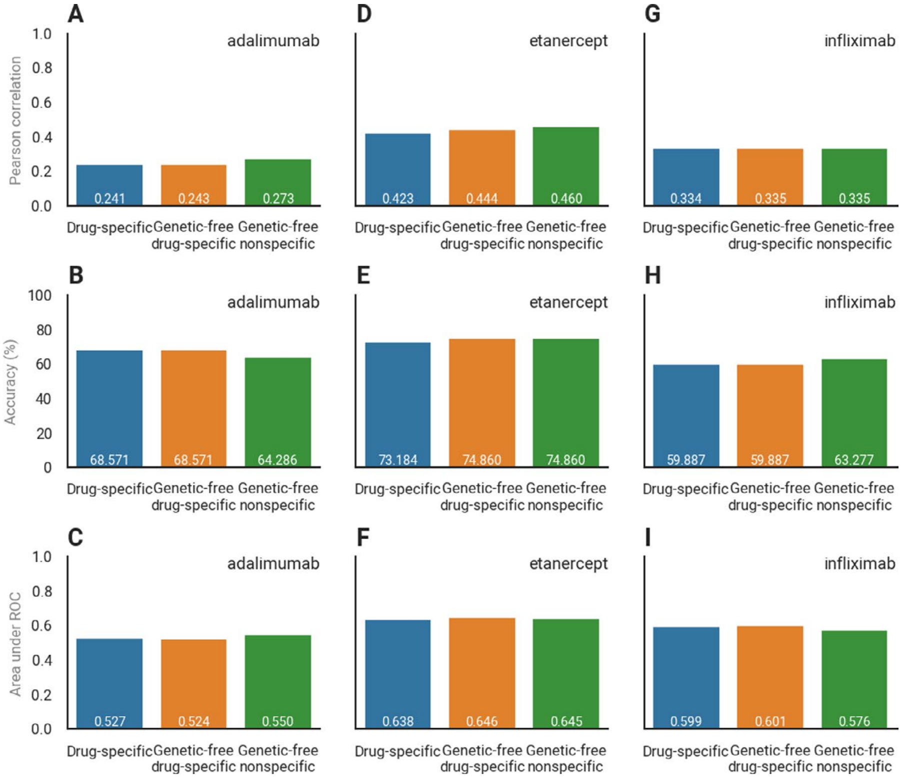

Figure 5. Evaluation of treatment-specific models and nonspecific models using the Comparative Effectiveness Registry to study Therapies for Arthritis and Inflammatory Conditions cohort data set. **A**, **D**, and **G**, Pearson's correlation coefficients between the observed change in Disease Activity Score in 28 joints (ΔDAS28) and predictions from tested regression methods for adalimumab, etanercept, and infliximab. **B**, **E**, and **H**, Classification accuracy ratio of predictions from tested responder versus nonresponder classification methods for adalimumab, etanercept, and infliximab. **C**, **F**, and **I**, Area under the receiver operating characteristic (ROC) curve of tested responder versus nonresponder classification methods for adalimumab, etanercept, and infliximab.

Table 2, http://onlinelibrary.wiley.com/doi/10.1002/art.41056/abstract). To assess these markers, we developed a baseline features–only model, a genetic features–only model, and an age/sex/methotrexate use model. These models were compared to the originally submitted GPR models, which incorporated all of these features. We performed both cross-validation tests using the training data set (Figure 4) and an independent test using the CERTAIN cohort data set (Figure 5). The results showed that while clinical information and genetic markers have relatively low predictive power themselves, they can improve the accuracy of the GPR model when the baseline DAS28 feature is added, especially for ΔDAS28 prediction. Our analysis of genetic markers showed a strong cohort association (Supplementary Figure 6, http://onlinelibrary.wiley.com/doi/10.1002/art.41056/abstract). Considering that the training cohorts were of European descent only and each cohort was from a specific geographic area, it is challenging to apply these biomarkers to other cohorts.

DISCUSSION

In this study, we demonstrated the state-of-the-art predictive power of the GPR model (39). The similarity modeling approach of GPR complements the ongoing development of precision medicine efforts in RA (40). The premise of similarity modeling has been widely used in social network analysis and other areas (41,42). It is considered to be effective in investigating heterogeneous data sets, which are commonly seen in cross-sectional studies (43). Additionally, the heterogeneity of diseases often obstructs explicit modeling of underlying distributions of individual features, which can be even more problematic when the sample population is small (44). A GPR model circumvents this issue by matching patients to those with similar conditions. The model developed in this study can predict which subpopulations will not respond to certain treatments, which can help physicians tailor treatments for individual patients based on their conditions.

The GPR model, as an interpretable method, has practical advantages in clinical application. Many sophisticated machine learning algorithms may make accurate predictions but lack interpretability for medical application (45–47). In contrast, GPR is a well-studied statistical model. The similarity modeling approach is intuitive, and its results are easy to interpret (48). In treatment response prediction, the kernel function allows for the identification of known subjects with similar conditions. While the kernel function for both GPR and SVM models provides information regarding the importance of genetic and clinical features, GPR also bases its prediction on the most similar individuals in the training data set. This allows physicians to inspect the conditions of known samples that have the highest weights in the GPR prediction. The additive design of our custom kernel function allows new features to be easily incorporated with reduced parameter tuning. This model can also estimate confidence intervals for its predictions, which can be useful for physicians.

However, in comparison to many linear methods that come with feature penalty, which allows for built-in feature selection, GPR does not have this benefit. As there are millions of genetic features, many genes are correlated with clinical outcomes by chance. Furthermore, because of the model differences, significance in linear model tests does not translate directly to accuracy improvement in a GPR model. To address these issues together, we chose a preselected set of genetic features in our model. The selection was based on cross-validation, with clinical features present in the training process. The genes chosen through this process were often by themselves not informative enough to predict the outcomes but may improve the performance of the clinical characteristics–only models. Recent advances in deep learning may aid in the future development of methodologies that directly connect genome sequences to phenotypes and treatment responses in RA by implementing one-dimensional convolutional neural networks that can be used to directly extract information from DNA sequences. With the increase in the amount of sequencing data, we foresee growth in this area in the near future.

Clinical and genetic heterogeneity may pose a major challenge for predicting anti-TNF responses. On one hand, clinical and demographic markers worked well across different cohorts. The cross-validation showed that although clinical markers themselves had predictive power, they also improved accuracy when added to the baseline DAS28–only model, regardless of cohort. Previous studies also identified several blood biomarkers that were informative about treatment responses and were validated across different cohorts (49,50). Genetic features used in our models have been reported to be immune-related, including insulin secretion (PDZD2), immunoresponse (CD84), and eicosanoids synthesis (PLA2G4A). Previous studies have shown ethnic differences in genetic markers for anti-TNF responses in the treatment of both RA (15,51) and other related autoimmune diseases (52,53). Our principal components analysis on the training data set showed that the genetic markers were associated with cohort information. We showed that genetic markers could not improve the prediction accuracy within the CERTAIN cohort data set as they did within the training data set. While both the training and the testing data in our study involved only patients of European descent, the variation across different geographic areas still obstructed the modeling of genetic markers. Accurate modeling of treatment responses would require a larger panel of genetic features covering multiple populations. We believe that extending the clinical feature panel to include blood markers or other clinical assays would be beneficial for cross-sectional predictions. On the other hand, genetic markers were found to be specific to certain populations in the context of sufficient genetic subtype modeling.

Compared to traditional trial-and-error methods, our model can help up to 40% of European-descent anti-TNF nonresponders avoid ineffective treatments. The model's performance is comparable to that of some published models that used additional biomarker data, whose AUCs ranged from 55% to ~74% using various testing sets (50). We would caution future users of this model that it was built upon data from European descendants only. Considering the heterogeneity of the anti-TNF responses among RA patients, we do not expect the model to perform similarly in other populations; the use of this model in other populations would require new patient data and separate feature selection.

In conclusion, we developed a GPR model to predict anti-TNF responses among RA patients and to identify nonresponders. The model interpretation shows promise in guiding treatment selection. While we showed that the clinical features described here are still the features most predictive of treatment response, the prediction model allows researchers to assess the contribution of genetic markers using existing clinical information across cohorts. In the future, various clinical markers may potentially be used for more accurate identification of nonresponder subpopulations that carry predictive biochemical traits (50,54). However, since this model was developed using data on European descendants only, transferring prediction models to other populations may be difficult (55). Further studies involving larger and more diverse populations will result in the development of more robust models to predict ΔDAS28.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. Guan had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study conception and design. Guan, Zhu.
Acquisition of data. Guan, Pappas, Kremer.
Analysis and interpretation of data. Guan, Zhang, Quang, Wang, Parker, Zhu.

ADDITIONAL DISCLOSURES

Author Zhang is an employee of Microsoft, Inc. Authors Pappas and Kremer are employees of Corrona, LLC.

REFERENCES

1. Geiler J, Buch M, McDermott MF. Anti-TNF treatment in rheumatoid arthritis. Curr Pharm Des 2011;17:3141–54.
2. Wijbrandts CA, Tak PP. Prediction of response to targeted treatment in rheumatoid arthritis. Mayo Clin Proc 2017;92:1129–43.
3. Lipsky PE, van der Heijde DM, St. Clair EW, Furst DE, Breedveld FC, Kalden JR, et al, for the Anti–Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. Infliximab and methotrexate in the treatment of rheumatoid arthritis. N Engl J Med 2000;343:1594–602.

4. Oliver J, Plant D, Webster AP, Barton A. Genetic and genomic markers of anti-TNF treatment response in rheumatoid arthritis. Biomark Med 2015;9:499–512.
5. Prajapati R, Plant D, Barton A. Genetic and genomic predictors of anti-TNF response. Pharmacogenomics 2011;12:1571–85.
6. De la Torre I, Valor L, Nieto JC, Hernández-Flórez D, Martinez L, Gonzalez CM, et al. Anti-TNF treatments in rheumatoid arthritis: economic impact of dosage modification. Expert Rev Pharmacoecon Outcomes Res 2013;13:407–14.
7. Roda G, Jharap B, Neeraj N, Colombel JF. Loss of response to anti-TNFs: definition, epidemiology, and management. Clin Transl Gastroenterol 2016;7:e135.
8. Aletaha D, Kapral T, Smolen JS. Toxicity profiles of traditional disease modifying antirheumatic drugs for rheumatoid arthritis. Ann Rheum Dis 2003;62:482–6.
9. Antoni C, Braun J. Side effects of anti-TNF therapy: current knowledge. Clin Exp Rheumatol 2002;20 Suppl 28:S152–7.
10. Burmester GR, Landewé R, Genovese MC, Friedman AW, Pfeifer ND, Varothai NA, et al. Adalimumab long-term safety: infections, vaccination response and pregnancy outcomes in patients with rheumatoid arthritis. Ann Rheum Dis 2017;76:414–7.
11. Hyrich KL, Watson KD, Silman AJ, Symmons DP, The BSR Biologics Register. Predictors of response to anti-TNF-α therapy among patients with rheumatoid arthritis: results from the British Society for Rheumatology Biologics Register. Rheumatology (Oxford) 2006;45:1558–65.
12. Atzeni F, Antivalle M, Pallavicini FB, Caporali R, Bazzani C, Gorla R, et al. Predicting response to anti-TNF treatment in rheumatoid arthritis patients. Autoimmun Rev 2009;8:431–7.
13. Umiċeviċ Mirkov M, Cui J, Vermeulen SH, Stahl EA, Toonen EJ, Makkinje RR, et al. Genome-wide association analysis of anti-TNF drug response in patients with rheumatoid arthritis. Ann Rheum Dis 2013;72:1375–81.
14. Cui J, Saevarsdottir S, Thomson B, Padyukov L, van der Helm-van Mil AH, Nititham J, et al. Rheumatoid arthritis risk allele PTPRC is also associated with response to anti–tumor necrosis factor α therapy. Arthritis Rheum 2010;62:1849–61.
15. Sode J, Vogel U, Bank S, Andersen PS, Thomsen MK, Hetland ML, et al. Anti-TNF treatment response in rheumatoid arthritis patients is associated with genetic variation in the NLRP3-inflammasome. PLoS One 2014;9:e100361.
16. Wu C, Wang S, Xian P, Yang L, Chen Y, Mo X. Effect of anti-TNF antibodies on clinical response in rheumatoid arthritis patients: a meta-analysis. Biomed Res Int 2016;2016:7185708.
17. Yamamoto K, Okada Y, Suzuki A, Kochi Y. Genetic studies of rheumatoid arthritis. Proc Jpn Acad Ser B Phys Biol Sci 2015;91:410–22.
18. Weyand CM, Klimiuk PA, Goronzy JJ. Heterogeneity of rheumatoid arthritis: from phenotypes to genotypes. Springer Semin Immunopathol 1998;20:5–22.
19. Plenge RM, Greenberg JD, Mangravite LM, Derry JM, Stahl EA, Coenen MJ, et al. Crowdsourcing genetic prediction of clinical utility in the Rheumatoid Arthritis Responder Challenge [letter]. Nat Genet 2013;45:468–9.
20. Sieberts SK, Zhu F, García-García J, Stahl E, Pratap A, Pandey G, et al. Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis. Nat Commun 2016;7:12460.
21. Prevoo ML, van ‘t Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight–joint counts: development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum 1995;38:44–8.
22. Van Gestel AM, Prevoo ML, van ’t Hof MA, van Rijswijk MH, van de Putte LB, van Riel PL. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis: comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism criteria. Arthritis Rheum 1996;39:34–40.
23. Liu C, Batliwalla F, Li W, Lee A, Roubenoff R, Beckman E, et al. Genome-wide association scan identifies candidate polymorphisms associated with differential response to anti-TNF treatment in rheumatoid arthritis. Mol Med 2008;14:575–81.
24. Allaart CF, Goekoop-Ruiterman YP, de Vries-Bouwstra JK, Breedveld FC, Dijkmans BA, FARR study group. Aiming at low disease activity in rheumatoid arthritis with initial combination therapy or initial monotherapy strategies: the BeSt study. Clin Exp Rheumatol 2006;24 Suppl 43:S77–82.
25. Bluett J, Morgan C, Thurston L, Plant D, Hyrich KL, Morgan AW, et al. Impact of inadequate adherence on response to subcutaneously administered anti-tumour necrosis factor drugs: results from the Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate cohort. Rheumatology (Oxford) 2015;54:494–9.
26. Plant D, Bowes J, Potter C, Hyrich KL, Morgan AW, Wilson AG, et al. Genome-wide association study of genetic predictors of anti–tumor necrosis factor treatment efficacy in rheumatoid arthritis identifies associations with polymorphisms at seven loci. Arthritis Rheum 2011;63:645–53.
27. Iannaccone CK, Lee YC, Cui J, Frits ML, Glass RJ, Plenge RM, et al. Using genetic and clinical data to understand response to disease-modifying anti-rheumatic drug therapy: data from the Brigham and Women’s Hospital Rheumatoid Arthritis Sequential Study. Rheumatology (Oxford) 2011;50:40–6.
28. Orellana C, Wedrén S, Källberg H, Holmqvist M, Karlson EW, Alfredsson L, et al. Parity and the risk of developing rheumatoid arthritis: results from the Swedish Epidemiological Investigation of Rheumatoid Arthritis study. Ann Rheum Dis 2014;73:752–5.
29. Bathon JM, Genovese MC. The Early Rheumatoid Arthritis (ERA) trial comparing the efficacy and safety of etanercept and methotrexate. Clin Exp Rheumatol 2003;21 Suppl 31:S195–7.
30. Coenen MJ, Enevold C, Barrera P, Schijvenaars MM, Toonen EJ, Scheffer H, et al. Genetic variants in Toll-like receptors are not associated with rheumatoid arthritis susceptibility or anti-tumour necrosis factor treatment outcome. PLoS One 2010;5:e14326.
31. Toonen EJ, Coenen MJ, Kievit W, Fransen J, Eijsbouts AM, Scheffer H, et al. The tumour necrosis factor receptor superfamily member 1b 676T>G polymorphism in relation to response to infliximab and adalimumab treatment and disease severity in rheumatoid arthritis. Ann Rheum Dis 2008;67:1174–7.
32. Miceli-Richard C, Comets E, Verstuyft C, Tamouza R, Loiseau P, Ravaud P, et al. A single tumour necrosis factor haplotype influences the response to adalimumab in rheumatoid arthritis. Ann Rheum Dis 2008;67:478–84.
33. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.
34. Kremer JM. The CORRONA database. Autoimmun Rev 2006;5:46–54.
35. Pappas DA, Kremer JM, Reed G, Greenberg JD, Curtis JR. Design characteristics of the CORRONA CERTAIN study: a comparative effectiveness study of biologic agents for rheumatoid arthritis patients. BMC Musculoskelet Disord 2014;15:113.
36. Aletaha D, Nell VP, Stamm T, Uffmann M, Pflugbeil S, Machold K, et al. Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. Arthritis Res Ther 2005;7:R796–806.

37. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods 2011;8:833–5.
38. Kim JH. Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. Comput Stat Data Anal 2009;53:3735–45.
39. Rasmussen CE, Williams CK. Gaussian processes for machine learning. Cambridge: MIT Press; 2006.
40. Kłak A, Paradowska-Gorycka A, Kwiatkowska B, Raciborski F. Personalized medicine in rheumatology. Reumatologia 2016;54:177–86.
41. Liu C, Liu J, Jiang Z. A multiobjective evolutionary algorithm based on similarity for community detection from signed social networks. IEEE Trans Cybern 2014;44:2274–87.
42. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS One 2011;6:e26752.
43. Huang Z, Zhang H, Boss J, Goutman SA, Mukherjee B, Dinov ID, et al, for the Pooled Resource Open-Access ALS Clinical Trials Consortium. Complete hazard ranking to analyze right-censored data: an ALS survival study. PLoS Comput Biol 2017;13:e1005887.
44. Rose NR, Mackay IR. The autoimmune diseases. 5th ed. Oxford: Academic Press; 2013.
45. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2006;2:59–77.
46. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015. p. 1721–30.
47. Che Z, Purushotham S, Khemani R, Liu Y. Interpretable deep models for ICU outcome prediction. AMIA Annu Symp Proc 2017;2016:371–80.
48. Caywood MS, Roberts DM, Colombe JB, Greenwald HS, Weiland MZ. Gaussian Process Regression for predictive but interpretable machine learning models: an example of predicting mental workload across tasks. Front Hum Neurosci 2017;10:647.
49. Cuchacovich M, Bueno D, Carvajal R, Bravo N, Aguillón JC, Catalán D, et al. Clinical parameters and biomarkers for anti-TNF treatment prognosis in rheumatoid arthritis patients. Clin Rheumatol 2014;33:1707–14.
50. Thomson TM, Lescarbeau RM, Drubin DA, Laifenfeld D, de Graaf D, Fryburg DA, et al. Blood-based identification of non-responders to anti-TNF therapy in rheumatoid arthritis. BMC Med Genomics 2015;8:26.
51. Honne K, Hallgrímsdóttir I, Wu C, Sebro R, Jewell NP, Sakurai T, et al. A longitudinal genome-wide association study of anti-tumor necrosis factor response among Japanese patients with rheumatoid arthritis. Arthritis Res Ther 2016;18:12.
52. Adshead R, Tahir H, Bubbear J, Donnelly S, Chau I. Ethnic differences in the response to anti-TNF in patients with ankylosing spondylitis [abstract]. Rheumatology (Oxford) 2015;54 Suppl:i133–4.
53. Liu J, Dong Z, Zhu Q, He D, Ma Y, Du A, et al. TNF-α promoter polymorphisms predict the response to etanercept more powerfully than that to infliximab/adalimumab in spondyloarthritis. Sci Rep 2016;6:32202.
54. Hueber W, Tomooka BH, Batliwalla F, Li W, Monach PA, Tibshirani RJ, et al. Blood autoantibody and cytokine profiles predict response to anti-tumor necrosis factor therapy in rheumatoid arthritis. Arthritis Res Ther 2009;11:R76.
55. Helliwell PS, Ibrahim G. Ethnic differences in responses to disease modifying drugs. Rheumatology (Oxford) 2003;42:1197–201.