Predicting Response to Tocilizumab Monotherapy in Rheumatoid Arthritis: A Real-World Data Analysis Using Machine Learning

Fredrik D Johansson (fredrik.johansson@chalmers.se)
Chalmers University of Technology: Chalmers tekniska hogskola
Jamie E Collins
Brigham and Women's Hospital
Vincent Yau
Genentech Inc
Hongshu Guan
Brigham and Women's Hospital
Seoyoung C Kim
Brigham and Women's Hospital
Elena Losina
Brigham and Women's Hospital
David Sontag
Massachusetts Institute of Technology
Jacklyn Stratton
Brigham and Women's Hospital
Huong Trinh
Genentech Inc
Jeffrey Greenberg
Corrona
Daniel H Solomon
Brigham and Women's Hospital

Research article

Keywords: Rheumatoid arthritis, disease-modifying anti-rheumatic drug, remission, prediction model, machine learning

Posted Date: September 24th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-79368/v1

Version of Record: A version of this preprint was published at The Journal of Rheumatology on May 1st, 2021. See the published version at https://doi.org/10.3899/jrheum.201626.

Abstract

Background

Tocilizumab (TCZ) had similar efficacy when used as monotherapy or in combination with other treatments for rheumatoid arthritis (RA) in randomized controlled trials (RCT). Recently, we derived a remission prediction score for TCZ monotherapy (TCZm) using RCT data. Herein, we describe external validation and several extensions of the prediction score using “real world data” (RWD).

Methods

We identified patients in Corrona-RA who used TCZm (n=453), matching the design and patients from four RCTs used in previous work (n=853). Patients were followed to determine remission status at 24 weeks. We compared the performance of remission prediction models in RWD, first based on variables determined in our prior work in RCTs, and then using an extended variable set, comparing logistic regression and random forest models. We included patients on other biologic DMARD monotherapies (bDMARDm) to improve prediction.

Results

The fraction of patients observed reaching remission on TCZm by their follow-up visit was 12% (n=53) in RWD vs 15% (n=127) in RCTs. Discrimination was good in RWD for the risk score developed in RCTS with AUROC of 0.70 (95% CI 0.64, 0.77). Fitting the same logistic regression model to all bDMARDm patients in the RWD improved the AUROC on TCZm patients to 0.73 (95% CI 0.64, 0.82). Extending the variable set and adding regularization further increased it to 0.77 (95% CI 0.68, 0.85).

Conclusion

The remission prediction scores, derived in RCTs, discriminated patients in RWD about as well as in RCTs. Discrimination was further improved by retraining models on RWD, including a larger variable set and learning from patients on similar therapies.

Significance And Innovation

- The observed fraction of patients reaching remission on monotherapy with tocilizumab by their follow-up visit at 6 months was 12% in real world data, very similar to the 15% observed in RCTs.
- Discrimination was as good in real world data for the risk score developed in RCTs with area under the receiver operative curve of 0.70 (95% CI 0.64, 0.77). Retraining and evaluating the risk model in RWD, extending the variable set and adding regularization further increased it to 0.77 (95% CI 0.68, 0.85).
- A remission prediction rule for tocilizumab monotherapy derived in RCTs performed well in real world data. The methodology used to derive the rule is generally applicable to other RA treatments.

Introduction

An expanding treatment armamentarium means more treatment options for patients with rheumatoid arthritis (RA), however clinicians face difficult decisions attempting to make evidence-based recommendations regarding which disease-modifying antirheumatic drug (DMARD) treatment will be most effective in a given patient. While the majority of patients with RA will find an effective treatment, not all do; many spend months trying medications that may not work for them.[1] Prior investigations have attempted to find biomarkers which can help personalize treatments, but most efforts have not produced useful results.[2] While an exhaustive comparison between all treatment options is a desirable goal, a natural first step is to identify and understand predictors of a single drug's success.

We recently examined clinical data from several randomized controlled trials (RCTs) and were able to derive and validate a prediction score for remission among patients using tocilizumab monotherapy (TCZm).[3] Monotherapy with TCZ has been found more effective than monotherapy with some targeted therapies for RA.[4] However, RCT data may not always replicate in typical practice with real world data.[5] These differences may derive from different patient populations, different treatment patterns, or other more subtle differences.[5]

Real-world data (RWD) offer important advantages over RCTs in that patients are more heterogeneous, with a greater variety of clinical characteristics as well as experience with other biologic and targeted DMARD treatments. We examined the performance of our original prediction score[3] for remission among patients using TCZm among patients in Corrona, a large RWD set from the US.[6] We employed various machine learning algorithms to take advantage of the copious data contained within Corrona.

Methods

Study Design

Our study sought to answer the following questions regarding remission in patients with RA using TCZm: A) to what extent do the findings of Collins et al[3] replicate in RWD? B.1) will expanding the set of remission predictors improve model fit? B.2) can data from patients on other therapies improve a predictive model for TCZm patients? C) what gains are there from applying non-parametric estimators of remission probability? We address these questions in sequence as described below.

Derivation and validation of original prediction score in RCT

As a baseline model, we used the remission model derived by Collins et al.[3] derived from patients on TCZm from two RCTs.[7,8] In the Collins study, twelve variables representing demographics, basic RA characteristics and treatment history were included in a logistic regression (LR) model based on three alternative criteria: an a priori baseline set of covariates, the model odds ratio (OR), and the model fit (expressed as the Akaike Information Criterion, AIC). In the current analyses, we restricted our comparison

to the set of variables determined by ORs, as this was deemed the most successful model, referred to here as LR-OR.

In the analysis by Collins et al.,[3] two RCTs were used for validation and two additional RCTs were used for validation.[4,7–9] In the current analyses, our focus is predictive discrimination in RWD. Hence, all four trials were used for derivation of models validated in the RWD. Access to the RCT and RWD data was granted following de-identification and IRB approval from the Partners Healthcare Human Studies. Variables from the four RCTs were harmonized by Collins et al. The less than 5% of subjects with missing values were removed from the study. Here, only TCZm patients were used for validation.

Question A. Evaluating baseline model in real-world data

Following the development of the initial LR-OR model in the RCTs, we evaluated the model using patients from the RWD dataset. This was pursued by using the parameter values fit in the RCTs, and also by refitting the parameters of the LR-OR model to RWD. Using these two complementary methods serves to estimate the extent to which the quality of the LR-OR model is affected by the cohort discrepancies between RCT and RWD.

Question B. Expanding the variable set and derivation population

The original variable set used in LR-OR was limited by the covariates collected in the four RCTs underlying its derivation.[4,7–9] In the Corrona RWD used in the current analyses, we had a greatly expanded feature set not available in RWD, including additional demographic information, comorbidities and treatment history. All models fit with this expanded set were trained and evaluated using only RWD. Adding covariates comes at a “statistical power price” however, since they contribute to increased variance if the cohort remains of fixed size. To address this, we used a regularized logistic regression model (LR-Reg) which penalizes models with many large coefficients.

To further reduce the variance of our model parameters, we evaluated models fit in an expanded population by including patients on monotherapy with *any* biologic DMARD, including TCZm, (bDMARDm). Our motivation for this is that variables that predict remission in RA are likely to be predictive for patients on different therapies. By including an indicator for TCZm therapy, this design choice greatly increased the cohort size while enabling the model to remain predictive for our cohort of interest. Validation was then pursued for the cohort including only patients using TCZm.

Question C. Applying ML algorithms in prediction of remission

A limitation of using logistic regression for predicting remission is that it models the log-odds of an event as a *linear* function of the covariates. It is standard practice in statistics to overcome this limitation by introducing features representing interactions (e.g., products of variables) or transformations (e.g., the square or logarithm) of variables. Effective transformations are often hard to know in advance and

exhaustive enumeration of all possibilities is not feasible, both for computational reasons and due to the exponential model size and problems with parameter variance it entails.

As an alternative, we used non-parametric *random forest* estimators—ensembles of tree-structured decision rules.[10] Random forests are universal function approximators capable of *discovering* meaningful interactions and transformations of variables while mitigating increased variance through the use of bootstrapping. A drawback of the random forest estimator is that it is often difficult to describe ensembles of learned decision rules concisely.[11] While the resulting predictions could serve as a computer-aided score, a more transparent description is often preferable. For this reason, we use random forests primarily to get an indication for how limited linear models are in this task.

Study Populations

Following Collins et al.,[3] we used four RCTs, the ACT-RAY and FUNCTION, ADACTA and AMBITION for derivation of our baseline model.[4,7–9]

Our RWD patient cohort was extracted from the Corrona RA registry—the largest prospective cohort study of RA in the world.[6] The registry comprises medical history including conditions, diagnoses, labs and treatments as well as demographic and lifestyle data. Records are collected through at regular visits through two questionnaires filled in by the patient and their physician, respectively. We used a version of the registry exported on February 4, 2018, containing 54,646 patients. This cohort was further restricted to patients who had been on bDMARDm therapy at some point recorded in the registry. The final dataset included records of visits from October 2001 to December 2017.

Patients in the Corrona RWD were deemed eligible for inclusion if they were on bDMARDm for a minimum of three months with at least one follow-up visit no later than nine months after initiation. Patients starting a bDMARD in combination with other DMARDs were included if the monotherapy with the target drug was started at most three months after target drug (not necessarily monotherapy) initiation; this occurred when the non bDMARD was stopped resulting in bDMARDm. A list of considered bDMARDs can be found in the **Appendix**.

Initially, to match the design of the RCTs, the Corrona cohort was restricted to patients older than 18 years on TCZm. Included patients had to be on monotherapy for at least 3 months and have a follow-up at most 9 months after initiation. Patients who started TCZ in combination with other DMARDs were included if they switched to monotherapy within 3 months of TCZ initiation. Follow-up duration was evaluated with respect to start of monotherapy. If patients were eligible at multiple time-points, only the first instance was included. For models fit to the full cohort of bDMARD subjects, an indicator variable for TCZ treatment was added. Finally, the most striking difference between the RWD and RCT cohorts was the average disease activity at baseline. For this reason, we evaluated models both for all RWD TCZm patients and for the subset of patients with baseline CDAI > 20 (see Tables 1,2).

Table 1

Baseline characteristics. N (%) or median (IQR). Variables in the original set across RWD (All, TCZm), RCTs, following imputation. RWD is also stratified by baseline CDAI > 20 to achieve a closer comparison to the RCTs. The extended variable set used in our analysis included variables representing: additional disease activity scores; history of cancer, hypertension, rheumatoid factor, joint erosions & deformity; comorbidities; prescriptions of NSAIDs and steroids; work status; education; general medical problems; physical disability; current and number of previous DMARDs.

Characteristic:	**All RWD (n = 3204)**	**TCZm (n = 452, RWD)**	**TCZm, CDAI > 20 (n = 240, RWD)**	**RCTs (n = 853)**
Age, years	57.0 (48.0, 66.0)	59.0 (49.0, 67.0)	59.0 (49.8, 67.0)	53.0 (44.0, 61.0)
Sex: Female	2439 (76.4%)	370 (82.0%)	194 (80.8%)	680 (79.7%)
Race: White	2932 (93.0%)	420 (94.4%)	227 (96.2%)	680 (79.7%)
BMI, kg/m^2	28.6 (24.9, 33.7)	28.7 (24.7, 33.7)	28.3 (24.0, 34.1)	26.5 (23.5, 30.5)
HAQ-DI	1.0 (0.4, 1.5)	1.2 (0.6, 1.6)	1.4 (1.0, 1.9)	1.6 (1.1, 2.0)
ESR, mm/hr	17.0 (8.0, 33.0)	16.0 (7.0, 36.0)	15.0 (6.5, 38.0)	40.0 (30.0, 60.0)
Hematocrit, %	40.0 (37.5, 43.0)	40.0 (37.4, 42.8)	39.8 (37.9, 42.9)	40.3 (40.1, 40.5)
Disease duration, years	8.0 (3.0, 16.0)	10.0 (5.0, 17.0)	10.0 (5.0, 18.0)	1.8 (0.5, 7.3)
Past DMARD/MTX:				
Both No	215 (6.7%)	14 (3.1%)	5 (2.1%)	291 (34.1%)
Both Yes	2669 (83.3%)	400 (88.5%)	216 (90.0%)	441 (51.7%)
DMARD Yes / MTX No	320 (10.0%)	38 (8.4%)	19 (7.9%)	121 (14.2%)
DMARD: TCZ	452 (14.1%)	452 (100.0%)	240 (100.0%)	853 (100.0%)
Baseline CDAI	17.4 (8.8, 28.0)	21.5 (12.0, 32.5)	32.0 (25.5, 40.3)	40.1 (30.7, 49.4)
Follow-up duration, weeks	32.7 (25.1, 52.7)	30.9 (25.0, 51.2)	28.1 (24.8, 42.1)	24.0 (24.0, 24.0)
Remission	563 (17.6%)	53 (11.7%)	13 (5.4%)	127 (14.9%)
Notes. RWD (Real-world data)				

Table 2

Discrimination in prediction of remission, measured by the area under the ROC curve (AUROC), for different models evaluated in the TCZm and TCZm (high CDAI) cohorts. Parentheses indicate 95% CIs. We compare models trained using the original variable set from Collins et al., (2019) and the extended feature set described in the Methods section. Additionally, we compare learning from only TCZm patients and learning from patients on any biologic DMARD monotherapy (bDMARDm). Cohort sizes (n = X) refer to the size of the respective validation set. AUROC near 0.5 is no better than random selection.

Model	Variables	AUROC, TCZm (n = 226)	AUROC, TCZm, CDAI > 20 (n = 120)
Model from Collins et al., (2019), trained on all RCT patients			
LR	Original	0.70 (0.64, 0.77)	0.56 (0.41–0.72)
Training on only RWD TCZm patients in derivation set			
LR	Original	0.68 (0.58, 0.78)	0.47 (0.23, 0.71)
LR	Extended	0.61 (0.51, 0.72)	0.52 (0.30, 0.75)
Random forest	Extended	0.74 (0.65, 0.83)	0.63 (0.45, 0.84)
LR-Reg	Extended	0.73 (0.64, 0.82)	0.67 (0.48, 0.85)
Training on all RWD bDMARDm patients in derivation set, evaluating on TCZm cohort			
LR	Original	0.73 (0.64, 0.82)	0.56 (0.34, 0.78)
LR	Extended	0.74 (0.65, 0.82)	0.67 (0.48, 0.86)
Random forest	Extended	0.76 (0.68, 0.84)	0.68 (0.50, 0.86)
LR-Reg	Extended	0.77 (0.68, 0.85)	0.72 (0.55, 0.89)

Notes. AUROC (Area under the receiver-operating characteristic curve) TCZ (Tocilizumab), TCZm (TCZ monotherapy), CDAI (Clinical Disease Activity Index), LR (Logistic Regression), LR-Reg (Regularized Logistic Regression), bDMARD (Biologic DMARD), RWD (Real-world data)

Study Outcome (RA Remission)

The primary outcome of interest was disease remission at 24 weeks following initiation of monotherapy, to match the outcome of the RCTs used by Collins et al.[3] Remission was defined by a Clinical Disease Activity Index (CDAI) < 2.8.[12] A benefit of the CDAI remission criteria is that it does not require access to a laboratory measurement and is thus widely available in RWD. In the Corrona RWD, remission status was evaluated at the follow-up visit closest to 24 weeks after start of monotherapy, but no sooner than 3 months or later than 9 months after initiation.

Potential Predictors

RCT Data

The variables used to in the baseline model derived from the RCTs were identical to Collins et al.[3] These included demographic variables (age, sex, geographic region, and body mass index (BMI)), as well as

several RA characteristics: baseline CDAI, disease duration, the Health Assessment Questionnaire (HAQ-DI), C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), and hematocrit. In addition, previous use of DMARDs (biologic or non-biologic) was recorded as MTX monotherapy, MTX plus another DMARD, any DMARD but not MTX, no DMARDs. The baseline variable set is listed in its entirety in Table 1.

Real-world Data

Most of the RWD was collected through a questionnaire filled out at each Corrona patient visit (typically once every six months) for each enrolled patient. Baseline features for the Corrona subjects were defined as the last recorded measurements taken *prior* to initiation of the target drug (TCZm or bDMARDm). In particular, if the target drug was prescribed for the first time between patient visits, the data of the last visit before prescription were used. To start, the variables of the baseline model LR-OR were extracted from the registry data underlying the RWD. In Corrona, the entire population resides in North America, so this predictor was omitted from models fit to the RWD.

In the case of missing data, variables were first forward-imputed based on visits prior to baseline. Remaining variables were either 0-imputed, in the case of indicators of certain comorbidities or lifestyle factors (e.g., smoking), or mean-imputed, in the case of continuous variables such as ESR. In the **Appendix**, we give statistics over the missingness of variables in the RWD before and after forward-imputation.

DMARD treatment status was determined using an algorithm previously used in the Corrona registry.[13] Patients were determined to be on a drug at a visit recorded in the registry if: i) the drug was prescribed to them at prior and next visit, ii) the patient reported using the drug at the current and next visit, and iii) the rheumatologist did stop the drug at the visit or reported initiating it at a later visit. Additionally, if the above holds for the previous and next visits and these were within 3 months of the current visit, the patient was determined to be on the drug at the current visit, as long as the rheumatologist did not report stopping the drug or initiating it at a later visit.

To investigate the predictive power of additional covariates, the baseline set was extended significantly using variables from the RWD. These variables included additional disease activity scores; history of cancer, hypertension, rheumatoid factor, joint erosions & deformity; additional comorbidities; prescriptions of non-steroidal anti-inflammatory drugs and glucocorticoids; work status; education; general medical problems; physical disability; current and number of previous DMARDs. A full description of this extended variable set (EXT) is provided in the **Appendix**.

Statistical Analyses

The envisioned clinical use-case for the developed risk scores is to aid in decision-making in treatment of *new* patients. Therefore, out-of-sample and out-of-distribution generalization is a primary concern. To address the former, as is customary, we use sample splitting of the cohort when training and validating models on RWD which are also evaluated on RWD. For a single experiment, the full sample was first split at random into a derivation set and a validation set, the former used for fitting model parameters and the

latter only for evaluation. The overall quality of each method was then computed as the average quality on the validation sets over a large number of repeated experiments. To assess out-of-distribution generalization, we evaluate the baseline model fit to the RCT cohorts on the RWD. The reverse (RWD to RCT) was not considered here as the extended variable set is not available in the RCTs.

The primary quality metric used was the area under the receiver-operating curve (AUROC). The AUROC assesses the extent to which models successfully *discriminate* (or rank) subjects in terms of their probability of remission. Standard errors in the AUROC were computed using the classical model of Hanley & McNeil.[14] A limitation of the AUROC is that it is not informative of the *calibration* of predicted probabilities—their absolute error with respect to the true probability. Typically, regularized classifiers—such as the ones considered in this work—are not trained to achieve perfect calibration, but minimize the expected out-of-sample classification error. However, given a well-discriminating model, calibration is often ensured or improved by applying, e.g., Platt scaling fit to held-out data.[15] We adopt this strategy here and assess calibration using standard methods.

Three different predictive models were fit to data and evaluated: logistic regression (LR), L1-regularized logistic regression (LR-Reg) and random forests. These estimators were chosen to illustrate the potential differences between a parametric (LR, LR-Reg) and a non-parametric model (random forests) and between regularized (LR-Reg) and unregularized models (LR). A significantly better fit for random forests compared to LR would point to the existence of important nonlinearities or covariate interactions in the target function. We describe our chosen models below.

Logistic Regression

Our baseline model was the (unregularized) logistic regression (LR) risk score developed by Collins et al.[3] with variables selected based on the odds ratio in the RCTs, while forcing the inclusion of sex, age and baseline CDAI into the model. To assess the value of including additional covariates from the RWD, we fit the same LR estimator to an extended variable set (EXT), as described in the previous section.

Regularized Logistic Regression with Ridge Penalty

Fitting a large number of model parameters to a small sample makes analyses susceptible to high variance.[16] Extending our variable set to include more covariates from the RWD runs this risk. However, under the assumption that only a few potential predictors have large predictive power, we may limit this using regularization.[16] Regularization trades off bias and variance by encouraging coefficients to be small in aggregate. Here, we used logistic regression with a ridge penalty, LR-Reg, which forces the sum of squared coefficients to be small. A regularization parameter controls the strength of the tradeoff. For small samples, a large penalty may be required to mitigate variance, but will introduce significant bias in coefficients. As our goal is primarily predictive capabilities, bias is of less concern. In larger samples, a smaller penalty is needed since variance is naturally controlled.

Random Forests

Random forests are ensembles of decision trees, each fit to a random subset of variables and subjects. The resulting estimator has a high capacity to discover interactions between variables and non-linear dependencies in the target variable, while estimating and mitigating variance by aggregating predictions of many classifiers.[10,17] They are the first choice of non-parametric estimator in many application areas. Like many other non-parametric estimators, random forests control sample variance (overfitting) by restricting the capacity of the model through a number of tuning parameters associated with the ensemble itself and with each tree. We describe these and the process for selecting them later in this section.

Weighting of Subjects in Extended Cohort

Extending the cohort to include patients on bDMARDs other than TCZ (see Question B) induces a distributional shift between the development (all BDMARDm) and evaluation cohorts (only TCZm). In particular, TCZm patients make up a fairly small proportion of the overall RWD cohort and would have limited impact on model fit in a standard model. If the remission probability, as a function of patient covariates, is different for different therapies, this could cause a bias in model fit that is disadvantageous when applied to TCZm patients specifically. To minimize this bias, we made use of inverse propensity re-weighting, as is standard practice for handling distributional shift between treatment groups or more generally between different populations. A logistic regression model was fit to estimate the propensity of patients in the RWD to receive TCZm treatment compared to receiving any bDMARDm therapy (including TCZm). This propensity was used to compute weights for samples in the training and validation sets which up-weight patients that had higher propensity to be put on TCZm. The weights were then uses to fit weighted logistic regression and weighted random forest models tailored to predicting remission for TCZm patients.

Model tuning and feature importance

The LR-Reg and random forest models have so-called tuning parameters, common in machine learning, which control the way that model coefficients are fit to data. In particular, these are used to influence the bias-variance tradeoff to improve out-of-sample performance. For LR-Reg, the only tuning parameter controls the strength of regularization—the degree of preference for smaller squared coefficients. Random forests have many potential parameters, the details of which vary between implementation. Here, we tuned the *maximum depth (1, 2, 5, 20, or no max)*, the *maximum number of features (1, 2, 5, 10, or no max)*, the *minimum samples per leaf (5, 10, 20, 50, 100)*, and the *splitting criterion* (Gini or entropy) used by each tree in the forest.[18]

Following standard practice, tuning parameters for random forests and LR-Reg were selected based on 3-fold cross-validation within the derivation set using the area under the ROC-curve (AUROC) as selection criterion. Following selection of these parameters, all models were fit to the entire derivation set. All experiments were implemented in Python. The LR models were fit using the *statsmodels*[18] package and LR-Reg, random forests with *scikit-learn.*[17]

For linear models fit to standardized variables, the magnitudes of regression coefficients are often interpreted as measuring the variables' importance. For tree and forest models, it is common to rank variables by their ability to discriminate between subjects with different outcomes when used as splitting nodes in the trees. Here, we measure this so-called feature importance using the *mean decrease in impurity* (MDI), which is the standard in scikit-learn.

Results

Patient Sample Characteristics

From the RCTs, a total of 853 subjects were enrolled in the TCZm arms and had complete data. Among these, 80% were female and 80% were white. At baseline, the mean CDAI was 40.1 and 52% of subjects had been treated previously with both MTX and another DMARD. In the RWD, out of 54,646 subjects, 3204 subjects were identified fitting our criteria for bDMARDm. 76% of these subjects were female and 93% white. The mean baseline CDAI was 17.4 and 83% were previously treated with both MTX and another DMARD. In the bDMARDm cohort, 452 were treated with TCZm.

Missingness at baseline in the RWD was low (< 2%) for variables in the original feature set, with the exception of disease duration (11%), ESR (30%) and HAQ-DI (25%). For the extended feature set, indicator variables representing certain past comorbidities, joint erosion, rheumatoid factor, smoking and previous pregnancy had high (> 30%) missingness. For evaluation of models fit to the RWD, the RWD was repeatedly split into a validation set, containing 50% (n = 226) of TCZm subjects and 20% (n = 550) of other bDMARDm subjects, and a derivation set containing remaining subjects.

Derivation and Validation of the Prediction Model

The full results of our evaluation of different sets of predictors (original/extended), models (LR, LR-Reg, random forest) and derivation sets (RCT, RWD TCZm, RWD bDMARDm) are presented in Table 2. Each combination is evaluated for the validation sets within two cohorts: the RWD TCZm population and the subset of these with CDAI > 20. Training models with the extended feature set on the RCT cohort was infeasible as many covariates were not available. For a closer comparison with RCTs, we report results also for the subset of 240 RWD subjects with baseline CDAI > 20. The cohorts were comparable on demographic characteristics (see Table 1). The calibration of the LR and random forest models, following Platt scaling, is illustrated in Fig. 1.

For all combinations of derivation and validation sets, the LR-Reg and random forest models trained on the extended feature set demonstrated larger AUROCs than LR models using the original feature set. Note that validation is done only on TCZm subjects, irrespective of derivation set. For example, when trained on all bDMARDm subjects, LR-Reg (Extended) achieved 0.77 (95% CI 0.68, 0.85) AUROC compared to 0.73 (95% CI 0.64, 0.82) for LR (Original), with numbers in parenthesis indicating the 95% CIs. This suggests that there are gains to be made in predictive performance from including additional comorbidity,

lifestyle and treatment variables in the risk score. Note, however, that the CIs overlap. We saw no advantage of using the random forest model over the regularized logistic regression model, in either setting, indicating that the remaining variance in the outcome is unlikely to be due to underutilized interactions or nonlinearities.

We found that expanding the derivation cohort to include non-TCZ bDMARDm patients improved the AUROC for both the TCZm and the TCZm high-CDAI cohorts (see **bottom of** Table 2). Compare, for example 0.73 (95% CI 0.64, 0.82) AUROC for LR (Original) trained on bDMARDm to 0.68 (95% CI 0.58, 0.78) for the same model fit to TCZm patients only. For LR-Reg, the AUROCs were 0.77 (95% CI 0.68, 0.85) and 0.73 (95% CI 0.64, 0.82) when fitting to bDMARDm and TCZm, respectively. Comparable gains were seen for the CDAI > 20 group, but due to its small validation set, the variance in these results are high. The original model, derived in the RCTs performed substantially worse in the high-CDAI cohort than in the full TCZm cohort, even though the criterion CDAI > 20 was meant to increase the similarity to the RCTs. However, as we can see in Table 1, significant differences in the cohorts remained. The results may be partially explained by higher variance in outcome for the high-CDAI cohort, after controlling for baseline disease severity.

For all models and derivation sets, different measures of disease severity (e.g., DAS28 and CDAI) at baseline were consistently highly predictive of remission. In Table 3, we list the features with highest estimated importance in the LR-Reg and random forest models trained on the extended feature set of the full bDMARDm cohort, ordered by feature importance (random forests) or coefficient magnitude (LR-Reg). The highest-ranked features are mostly unsurprising: the majority pertain to measures of disease severity either explicitly (CDAI, MDAS, global assessment) or implicitly (larger number of previous DMARDs). For the LR-Reg model, education up to high-school or less was associated with lower chances of remission. The remission rate in this group (13%) was smaller than for subjects with more than high-school education (20%). The average reduction in CDAI between baseline and follow-up was almost identical for the two groups (-5.7 and − 5.8).

Table 3
Features with highest estimated importance measured by the mean decrease in impurity (MDI) for random forests and the magnitude of coefficients ιβι for LR-Reg.

Random Forest (Extended)	MDI	LR-Reg (Extended)	β (95% CI)
CDAI	0.10	In remission at baseline	0.30 (0.14, 0.39)
DAS	0.09	DAS	-0.30 (-0.38, -0.14)
MD Global Assessment	0.07	Work status: Disabled	-0.29 (-0.46, -0.25)
HAQ-DI	0.06	Past steroid prescription	-0.27 (-0.36, -0.07)
Remission at baseline	0.05	Education: High-school or less	-0.25 (-0.39, -0.15)
Trouble dressing self	0.04	Sex: Female	-0.25 (-0.36, -0.11)
Num. past DMARDs	0.04	Past DMARD/MTX: Both Yes	-0.25 (-0.36, -0.14)

Notes. LR-Reg (Regularized Logistic Regression), Extended (Extended variable set), CDAI (Clinical Disease Activity Index), DAS (Disease Activity Score), HS (High-school), HAQ-DI (Health Assessment Questionnaire without Disability Index).

Confidence intervals for LR-Reg coefficients were computed using the empirical bootstrap over the derivation set. This method was chosen due to the inclusion of regularization and sample weighting in the procedure. Subjects were resampled with replacement and the propensity and outcome models were fit to each bootstrap sample. The stated results are for the best tuning parameters from the experiment presented in Table 2.

Discussion

Machine learning applied to real-world data may offer new opportunities to better define the course of disease and to identify better treatment strategies. In RA, the expanded treatment options present a challenge for clinicians and patience, as predictors for response to specific treatments are lacking. In prior work, we used RCT data to derive and validate predictors of remission among patients initiating TCZm.[3] In the current study, we tested this prediction rule using RWD and attempted to refine the prediction rule using machine learning. We found that the original prediction rule held up well in RWD from Corrona, despite notable differences between the RCT and RWD populations. This and the fact that the original rule contains only commonly available variables points toward the feasibility of implementing these rules in clinical practice.

An expanded number of predictive variables were identified using machine learning algorithms which led to small gains in performance. However, for the features and number of samples available, we saw no gains in using a random forest model over logistic regression. This may be explained either by a lack of bias—that the conditional probability of remission is well approximated by the logistic model—or by noise and variance—that the available observations are insufficient to learn a better approximation without further assumptions. A larger (or broader) study, including comparison of a larger set of machine learning models, could help decide which of these explanations dominates the behavior.

The implications of this work are several. First, we examined the validity of prediction models derived and validated in trials in RWD. RCTs, while appropriate for estimating average treatment effects on the selected cohort, are limited in their generalizability to a broader population. RWD offers an insight into how patients are treated in the healthcare system, and what their outcomes are in the absence of strict inclusion criteria and potential experiment effects. In this work, for example, we found that the RCT and RWD cohorts differ substantially in terms of disease duration and severity, as well as treatment history. For example, the FUNCTION trial enrolled subjects that were MTX naïve with short disease duration,[4] while the other RCTs enrolled subjects that showed inadequate response to MTX.[7–9] Despite this, the discrimination between patients was good for the transferred model in terms of AUROC. This indicates that a) good predictors in the RCTs are good predictors in the RWD, and b) the variables observed for at baseline appropriately controlled for the cohort differences. To further validate the models we derived from the RWD, we may consider how well RWD generalizes to an RCT cohort. We refrained from this here as the RCT data did not contain the extended variable set used in the RWD analysis, nor are the RCT cohorts as representative as the RWD of current treatment practices.

Second, we developed and validated in RWD several prediction models for remission with TCZm. As noted above, TCZ when given as monotherapy seems to be more effective than other bDMARDs as monotherapy.[4] The variables identified in our prediction rule were disabled working status, educational attainment, prior DMARDs, and baseline CDAI. These variables are not surprising, but they have never been put together in one prediction rule that may have utility in the clinic. They also point out the value of several non-clinical variables. While we anticipate further refining this rule and testing it among other bDMARDs, it may be that future iterations of this rule could be programmed in an electronic medical record and help clinicians and patients identify therapy likely to be effective in patients with a given set of characteristics.

Finally, this set of analyses used a robust RWD dataset, Corrona, with an expanded set of variables. Because of the presence of many potentially correlated variables, we used several machine learning algorithms to analyze these data. In high-dimensional settings such as these, machine learning may be used together with sample splitting to discover models with reduced predictive variance at the cost of a small increase in bias.[16] This is appropriate particularly in applications where prediction and out-of-sample generalization is the goal, rather than parameter identification. The benefits of machine learning are smaller when domain knowledge is strong enough to identify a successful model without the need to search over a large set of variables. Additionally, in some cases, a model with slightly lower predictive accuracy may be preferred if it is easier to interpret, explain or communicate.[19]

Strengths of the current analyses include the validation of a previously derived and validated algorithm using RWD as an external validation dataset. However, several limitations should be noted. The work needs to be expanded to consider the prediction rule across other bDMARD; this is planned future work. We had significant rates of missing data in the RWD; this is typical but likely has some impact on model fit. In particular, different imputation strategies could be considered. Corrona encompasses patients from

North America only; while more generalizable and much larger than many RCTs, this is a limitation. The further generalizability of the prediction models developed here from Corrona to other RWD should be explored further.

Conclusion

In conclusion, we were able to test a prediction rule for remission with TCZm among patients with RA in RWD and found that it worked well. Additional variables enhanced the prediction rule further. Moreover, using data from other bDMARDs allowed us to improve the model fit. We encourage other investigators to derive and validate prediction models for RA treatment across RCTs and RWD. Machine learning algorithms may play important roles in optimizing prediction rules.

Declarations

- Ethics approval and consent to participate: All patients in the study database gave written informed consent.
- Consent for publication: Not applicable.
- Availability of data and material: The data used herein are property of Corrona. All qualified investigators are free to contact Corrona and inquire about use of the data.
- Competing interests: DHS receives salary support from research grants to Brigham and Women's Hospital from Abbvie, Amgen, Corrona, Janssen, Pfizer, and Roche/Genentech. He also serves on the editorial board of Arthritis & Rheumatology and on the FDA Arthritis Advisory Committee. FDJ receives salary support from research grants to Chalmers University of Technology from AstraZeneca. SCK received research grants to Brigham and Women's Hospital from AbbVie, Pfizer, Roche, and Bristol-Myers Squibb for unrelated studies. JEC is a consultant to Boston Imaging Core Labs and serves as Associate Editor for Statistics at Osteoarthritis and Cartilage.
- Funding: Roche/Genentech provided a research contract to Brigham and Women's Hospital.
- Authors' contributions

  - Conception: All

  - Data acquisition: Greenberg

  - Analysis: Johansson, Collins, Yau, Guan, Kim, Losina, Sontag, Solomon

  - Drafting manuscript: Johansson, Solomon

  - Revising manuscript: All
- Acknowledgements: Not applicable.

References

1. Aletaha D, Smolen JS. Diagnosis and Management of Rheumatoid Arthritis: A Review. *JAMA.* 2018;320(13):1360-1372.
2. Boire G, Allard-Chamard H. The 4-H of Biomarkers in Arthritis: A Lot of Help, Occasional Harm, Some Hype, Increasing Hope. *J Rheumatol.* 2019;46(7):758-763.
3. Collins JE, Johansson FD, Gale S, et al. Predicting Remission Among Patients With Rheumatoid Arthritis Starting Tocilizumab Monotherapy: Model Derivation and Remission Score Development. *ACR Open Rheumatol.* 2020;2(2):65-73.
4. Gabay C, Emery P, van Vollenhoven R, et al. Tocilizumab monotherapy versus adalimumab monotherapy for treatment of rheumatoid arthritis (ADACTA): a randomised, double-blind, controlled phase 4 trial. *Lancet.* 2013;381(9877):1541-1550.
5. Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? *Clin Pharmacol Ther.* 2017;102(6):924-933.
6. Kremer J. The CORRONA database. *Ann Rheum Dis.* 2005;64 Suppl 4:iv37-41.
7. Dougados M, Kissel K, Conaghan PG, et al. Clinical, radiographic and immunogenic effects after 1 year of tocilizumab-based treatment strategies in rheumatoid arthritis: the ACT-RAY study. *Ann Rheum Dis.* 2014;73(5):803-809.
8. Burmester GR, Rigby WF, van Vollenhoven RF, et al. Tocilizumab in early progressive rheumatoid arthritis: FUNCTION, a randomised controlled trial. *Ann Rheum Dis.* 2016;75(6):1081-1091.
9. Jones G, Sebba A, Gu J, et al. Comparison of tocilizumab monotherapy versus methotrexate monotherapy in patients with moderate to severe rheumatoid arthritis: the AMBITION study. *Ann Rheum Dis.* 2010;69(1):88-96.
10. Breiman L. Random forests. *Machine Learning.* 2001;45(1):5-32.
11. Friedman JH, Popescu BE. Predictive learning via rule ensembles. *The Annals of Applied Statistics.* 2008;2(3):916-954.
12. Aletaha D, Smolen J. The Simplified Disease Activity Index (SDAI) and the Clinical Disease Activity Index (CDAI): a review of their usefulness and validity in rheumatoid arthritis. *Clin Exp Rheumatol.* 2005;23(5 Suppl 39):S100-108.
13. Greenberg JD, Kremer JM, Curtis JR, et al. Tumour necrosis factor antagonist use and associated risk reduction of cardiovascular events among patients with rheumatoid arthritis. *Ann Rheum Dis.* 2011;70(4):576-582.
14. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology.* 1983;148(3):839-843.
15. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers.* 1999;10(3):61-74.
16. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media; 2009.

17. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research.* 2011;12(October):2825-2830.
18. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. Paper presented at: Proceedings of the 9th Python in Science Conference2010.
19. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence.* 2019;1(5):206-215.

Appendix

List of considered bDMARDs (as used in the Corrona registry)

Tocilizumab, Adalimumab, Certolizumab pegol, Etanercept, Golimumab, Infliximab, Abatacept, Tofacitinib, Baricitinib, Sarilumab, Rituximab

Extended Variable Set

The full set of variables extracted from the RWD are listed below.

Original set

Age, baseline CDAI, disease duration, sex, past DMARD use (MTX/other/neither), BMI, HAQ-DI, ESR, Hematocrit

Additional treatment variables

TCZ dose, steroid prescription, prednisone dose, NSAIDs (aspirin, celecoxib, ibuprofen, naproxen, diclofenac, other) , number of past DMARDs, indicator for current bDMARD at baseline

Additional disease severity markers / labs

MDAS, MD Global assessment, rheumatoid factor positive ever, CCP positive ever, joint erosions ever, joint narrowing ever, joint deformation ever, CRP

Comorbidities / medical history

History of cardiovascular disease, history of diabetes, history of hypertension, history of cancer, pregnant ever, Sjögren's disease, depression,

Lifestyle, education, work

Work status (working or student disabled, retired), education (college, high-school or less, other), smoking currently, exercise

Medical or mobility problems

Trouble bending down, trouble dressing self, trouble turning faucets, back pain, constipation, cough, depression, dizziness, dry eyes, dry mouth, headaches, heartburn, hair loss, memory problems, muscle pain, muscle weakening, nausea, numbness/tingling, shortness of breath, sleeping, thinking problems, unusual fatigue, weight gain

Figures

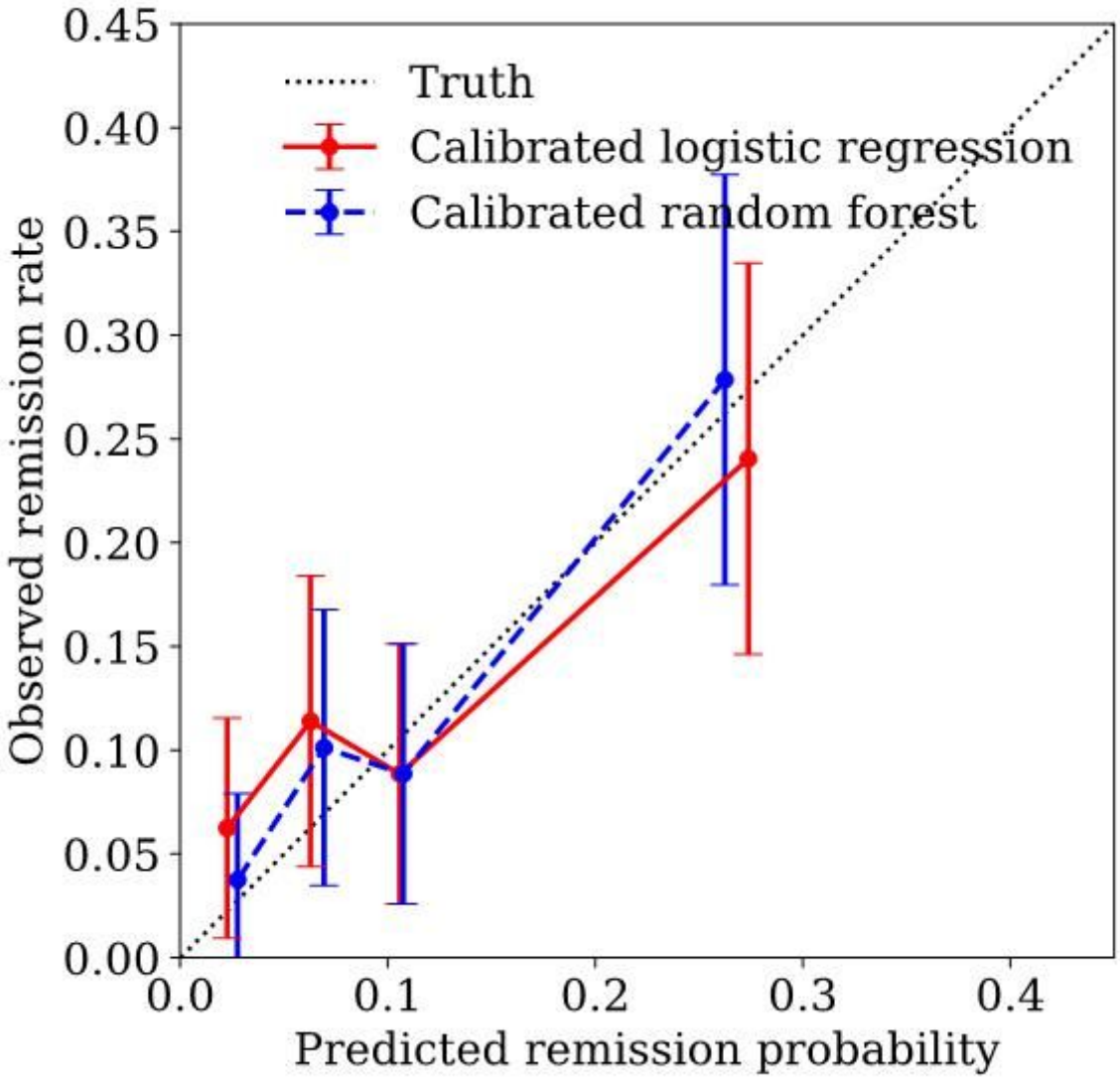

Figure 1

Calibration of logistic regression (LR) and random forest models trained on all bDMARDs in RWD using the extended feature set and evaluated on held-out TCZm patients from the RWD. The predictions of each model have been adjusted using Platt scaling. Calibration is assessed in the four quartiles of predicted remission probability. We note that the majority of patients (75%) have a predicted probability of remission around or below 0.1.